Proceedings of ICFTBA 2026 Symposium: Digital Government, Finance, Accountability, and Sustainable Economic Development
DOI: 10.54254/2754-1169/2026.AB39208

Forecasting Home Values in Los Angeles: Research Design and Empirical Results

Zheyan Yu

Pasadena City College, Pasadena, USA
zyu43@go.pasadena.edu

Abstract. This study examines how forecast accuracy varies across horizons for 85 Los Angeles County cities, the county, and California. The data are monthly Zillow Home Value Index values from January 2000 through August 2026. Holt's linear trend, damped exponential smoothing (ETS), autoregressive integrated moving average (ARIMA), and an equally weighted ARIMA and ETS combination are compared with naïve and drift benchmarks. Specifications are selected using data through 2016. Expanding window forecasts are evaluated one, three, six, and twelve months ahead over a 2017–2019 preliminary evaluation and a test period from January 2020 through August 2026. Errors are measured in dollars within locations and by mean absolute scaled error (MASE) across cities. At one month, ARIMA reduces MAE relative to drift by a median of 71.1% across cities. Its advantage declines as the horizon increases. At twelve months, drift has the lowest median error across cities, and ARIMA's MAE is a median of 9.8% higher. Drift also leads in every price and volatility tercile at this horizon, but rankings vary across subperiods. Residual correlation and convergence failures limit claims about ARIMA's fit. Accuracy estimates are retrospective because the data come from a single revised vintage. Naïve and drift benchmarks should remain part of model comparisons at every horizon.

Keywords: home value forecasting, Zillow Home Value Index, ARIMA, exponential smoothing, Los Angeles County

1. Introduction

Buyers, sellers, renters, and housing planners in Los Angeles use expectations about home values when considering timing, budgets, and planning decisions. A county average gives an overall picture but can hide differences in prices and market conditions across cities. This paper compares simple time series methods for cities in Los Angeles County at horizons of one, three, six, and twelve months. The question is which methods produce the lowest errors, rather than why home values change.

Some U.S. states have more predictable home prices than others. Rapach and Strauss [1] documented this variation, and Bork and Møller [2] compared forecasting models for all 50 states. Later studies moved to the city level outside the United States, with neural networks for 100 Chinese cities [3] and Gaussian-process regression for ten [4]. Closer to the present setting, Plakandaras et al. [5] forecast county-level prices in the 56 most populated U.S. metropolitan areas with a machine-

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.AB39208

learning estimator that allows for cross-sectional heterogeneity and dependence. None of these studies looks at cities inside Los Angeles County, and with different locations, models, and evaluation designs, their results do not carry over directly. Comparing methods across cities also requires holding the sample dates, horizons, and error measures fixed for every city; otherwise a gap between two methods may only reflect how each was scored.

This paper applies one design to all 85 cities in Los Angeles County and scores every model at the same four horizons, with Zillow's California and county series as regional context.

In this sample ARIMA has the lower median city error up to six months, but the gap narrows as the horizon lengthens. At twelve months, the simple drift benchmark has the lower full-test city median. Section 2 sets out the data and forecasting design; Sections 3 to 6 cover diagnostics, results, group and period checks, and limitations in turn.

2. Research design

Model specifications and the ARIMA order ranking are selected using data through 2016, before any forecast in the preliminary evaluation or test period is scored.

2.1. Data and sample

The data are Zillow's all-homes, middle-tier Zillow Home Value Index (ZHVI), a monthly, smoothed, and seasonally adjusted series [6]. It gives the typical value of homes between the 35th and 65th percentiles, not the mean price of homes sold in a given month. The city, county, and state CSV snapshots used here are dated September 29, 2026. Matching city records to the county's list of incorporated cities [7] on both the county and state identifiers gives 85 of the county's 88 cities. All 85 have complete monthly data from January 2000 through August 2026, or 320 observations per city. Industry and Vernon have no row in the city file and Westlake Village is coded to Ventura County, so these three are dropped, with nothing imputed in their place. The state and county figures come directly from Zillow's own geographic series; they are not built up from the 85 cities. Zillow builds the current ZHVI from its neural Zestimate only from 2012 onward. Values before 2012 come from its previous index, joined with an adjustment that reconciles the two seasonal patterns [8]. The January 2000–December 2016 training sample therefore spans this change in construction, while the pre-test and test periods come after it.

The three source files are kept locally as zillow_city_zhvi.csv, zillow_county_zhvi.csv, and zillow_state_zhvi.csv; they come from the Zillow housing-data page cited as [6].

The exercise is pseudo-out-of-sample: at each historical forecast origin the September 2026 vintage is cut at that date, so no later observation enters the fit. Values in that vintage may have been revised since they were first published. Zillow's release lags are not modeled, since each date refers to the month a value describes. The results therefore show how well the models would have predicted the retained ZHVI series in hindsight. Table 1 lists the sample and evaluation windows.

Table 1. Data split and evaluation windows

Component	Period	Purpose
Initial training	Jan. 2000–Dec. 2016 (204 months)	Rank ARIMA candidate orders and fit the initial models
Pre-test evaluation	Jan. 2017–Dec. 2019	Describe pre-test accuracy; do not select ARIMA orders

DOI: 10.54254/2754-1169/2026.AB39208

Table 1. (continued)

Pseudo-out-of-sample test	Jan. 2020–Aug. 2026 (80 target months)	Evaluate matched target-date errors
Geographic units	California, Los Angeles County, 85 cities	Compare state, county, and city forecasts

Note. The pre-test period has 36, 34, 31, and 25 evaluable target months at the 1-, 3-, 6-, and 12-month horizons; the test period has 80 at every horizon. Within a horizon, all models share the same origin–target pairs. Once a pre-test or test month has passed, later origins can use it, but no origin uses data from after its own date.

2.2. Forecasting models and horizons

The four main models are Holt's undamped linear trend, damped-trend exponential smoothing without seasonality (labeled ETS), ARIMA with drift, and an equal-weight combination of ARIMA and ETS. The naïve forecast and a random walk with drift serve as benchmarks and are left out of the rankings of the main models. Holt and ETS are fitted to log ZHVI with an additive trend, estimated initial states, and optimized smoothing parameters. They differ only in the damping parameter that ETS also estimates. ETS here means this fixed damped additive-trend specification, fitted with the ExponentialSmoothing class in statsmodels; the ETSModel class and automatic ETS model selection are not used [9, 10]. Holt uses the Holt class. Both are refitted at every forecast origin on the expanding sample, so their smoothing parameters and initial states are re-estimated every month. Forecasts from Holt, ETS, and ARIMA are returned to dollars by exponentiating after adding half the one-step residual variance of the model fitted at that origin. This variance is computed without the first twelve residuals. Forecast variance may differ by horizon, but the same correction is used at every horizon. The combination averages the ARIMA and ETS dollar forecasts [11].

For ARIMA, d is fixed at 1, and p and q each range from 0 to 2. A linear time trend in log levels supplies the drift after differencing. Each candidate order is fitted using up to 300 optimizer iterations under the software's stationarity and invertibility constraints. Converged candidates are ranked by the Akaike information criterion (AIC), and the Bayesian information criterion (BIC) is retained as a diagnostic. The convergence failures reported in Section 3.2 motivate a fallback rule at each forecast origin: the order ranked first is fitted using up to 200 iterations, then retried using up to 500 if needed. If neither attempt converges, the next order in the ranking is tried in the same way. Parameters are estimated again each month, but the ranking remains fixed after training.

At each forecast origin t, each fitted model produces a path through t+12. Forecasts for one, three, six, and twelve months ahead are drawn from this path rather than from separate direct regressions. ARIMA generates the path recursively, using its own predictions in subsequent steps. For example, the forecast for January 2020 made twelve months ahead has its origin in January 2019. Both benchmarks are computed in dollars: the naïve method carries the last observed value forward, and drift extends that value using the average monthly change from the first observation through t.

2.3. Evaluation rules

Within a city, dollar errors are summarized by root mean squared error (RMSE) and mean absolute error (MAE), with mean absolute percentage error (MAPE) added in Table 7. Across cities the main measure is a mean absolute scaled error (MASE) adapted from Hyndman and Koehler [12]. It divides test MAE by the city's mean absolute one-step naïve error over January 2000–December

DOI: 10.54254/2754-1169/2026.AB39208

2019 (through 2016 for the pre-test period). These fixed denominators play no part in model fitting, but because some longer-horizon test origins precede December 2019, the test scale could not have been known in real time. MASE is relative to typical historical monthly changes, so it is not a percentage of home value, and a lower MASE need not mean a smaller dollar loss.

A MASE below one says only that the forecast MAE is smaller than the historical one-step scaling error. Whether a model actually beats the naïve forecast at the same horizon is checked on matched test observations with the out-of-sample R^2. This statistic is one minus the ratio of the model's sum of squared forecast errors to the naïve sum.

The best performer for each city and horizon is the model with the lowest test MASE among Holt, ETS, ARIMA, and the combination. The gap equals the MASE of the model ranked second divided by the lowest MASE, minus one. Gaps below 2% are treated as practical ties so that small differences are not presented as clear wins. This cutoff is descriptive and is not based on a significance test or a financial threshold. The two benchmarks are compared separately with the main models.

The group comparisons examine whether accuracy varies with local price levels and historical volatility. Cities are divided into terciles separately for December 2019 ZHVI and the standard deviation of monthly log growth from 2000 to 2019, using the qcut function in pandas. The groups summarize test outcomes and never enter model fitting. At longer horizons, their defining measures may include observations dated after some forecast origins.

ARIMA and drift are compared formally with the Diebold–Mariano (DM) test [13]. Because it was chosen after earlier results had been seen, its twelve comparisons (three geographic levels by four horizons) are treated as exploratory. Loss is absolute error scaled by each unit's fixed historical one-step naïve MAE. For the city-level test, the losses are averaged with equal weight over the 85 cities in each target month. The original statistic uses a Bartlett long-run variance estimate [14] up to lag $h-1$, averaging the lag-k covariance over $80-k$ pairs. Its two-sided p-values come from a t distribution with 79 degrees of freedom. Harvey et al. [15] pair this reference distribution with a finite-sample correction to the statistic; only the t reference is used here, and no other finite-sample correction is made. A Holm adjustment [16] covers the twelve tests.

A further post hoc check reruns the same contrasts with Bartlett HAC bandwidths of 12 and 24 months. It uses a common denominator of 80 in the covariance estimates and the same t(79) approximation, again without a finite-sample correction. Holm adjustment is redone within each bandwidth, and results for both are reported; neither bandwidth was picked for its p-values. The 95% intervals given below are approximate and pointwise.

3. Data characteristics and model diagnostics

Price levels differ widely across the sample, and the evidence on stationarity is mixed.

3.1. Descriptive statistics

In August 2026 the ZHVI stood at $929,572 for Los Angeles City, $873,008 for the county, and $764,158 for the state (Table 2). Figure 1 rebases every series to 100 in January 2000 so that growth can be compared without these level differences.

DOI: 10.54254/2754-1169/2026.AB39208

Table 2. Descriptive statistics for monthly ZHVI (US dollars)

Unit	Mean	SD	Median	Minimum	Maximum	Aug. 2026
California	455,185	174,596	426,434	188,458	781,759	764,158
LA County	524,387	197,915	485,012	210,218	896,373	873,008
LA City	566,323	220,890	519,175	220,653	971,810	929,572
Median of 85 cities	523,373	195,950	493,979	216,238	893,549	879,254

Note. Each entry in the last row is the median, across the 85 cities, of that city-level statistic. Every unit has 320 monthly observations. SD is the standard deviation of the ZHVI level over the whole period, which is different from the volatility of monthly growth used to form groups in Section 5.

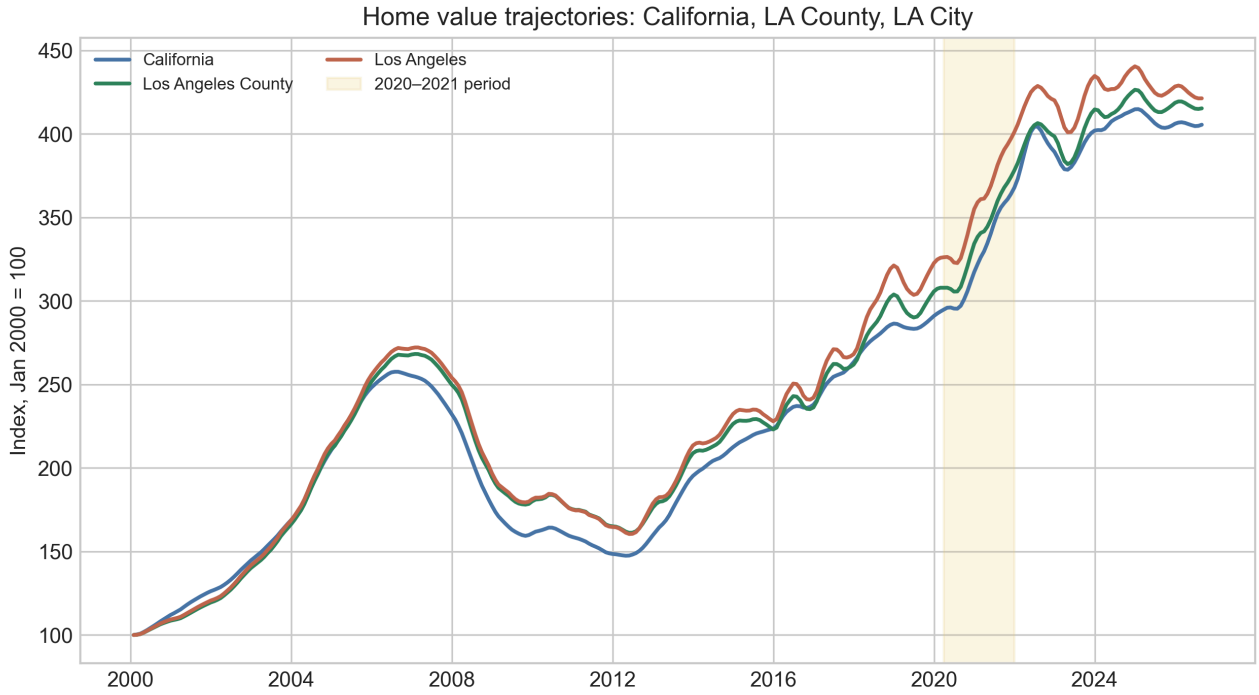


Figure 1. Normalized ZHVI, January 2000 = 100; shading marks March 2020–December 2021

3.2. Stationarity and ARIMA diagnostics

The augmented Dickey–Fuller (ADF) test [17, 18] takes a unit root as its null hypothesis and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test [19] takes stationarity. Applied to the initial January 2000–December 2016 sample (Table 3), both point to nonstationarity in log levels. ADF rejects a unit root for only 5 of the 87 series, while KPSS with a trend rejects stationarity for all 87. Differencing does not settle the matter, since ADF then rejects for none of the series and KPSS still rejects level stationarity for 30. An ADF test that rejects nothing after differencing is unusual, so it was rerun on the first difference of log ZHVI with a constant and AIC-selected lags. The rerun reproduced all 87 results, and a separate regression reconstruction matched the ADF statistics. The selected lags run from 7 to 15 months, and with fixed lags the number of rejections changes with lag length. The tests therefore cannot establish that every differenced series is stationary. The models nonetheless use d = 1 everywhere, to keep a single specification across regions.

DOI: 10.54254/2754-1169/2026.AB39208

Table 3. Stationarity tests in the initial training sample (87 series)

Test and null hypothesis	Log level: rejections at 5%	First difference: rejections at 5%
ADF: unit root	5/87	0/87
KPSS: trend/level stationarity	87/87	30/87

Note. The initial sample has 204 monthly levels and 203 first differences. ADF includes a constant and chooses among up to fifteen lags by AIC. KPSS uses a trend for log levels and a constant for first differences. Its bandwidth comes from the data-dependent rule of Hobijn et al. [20] (statsmodels nlags='auto'), which selects 9 lags for every log-level series and 8 or 9 for first differences. The two tests therefore differ in their deterministic terms.

Figure 2 shows the autocorrelation function (ACF) and partial autocorrelation function (PACF) for monthly log growth in Los Angeles City. The PACF is computed for lags 1 through 24 using Yule–Walker estimates without adjustment for sample size (statsmodels method 'ywm'). The dashed lines at approximately ±0.138 show pointwise bounds for a white noise series containing 203 observations. The ACF declines slowly and remains above the upper bound for more than 20 lags. The PACF has a large positive spike at lag 1 and a large negative spike at lag 2. This pattern supports including autoregressive terms. A specific ARIMA order and the stationarity of the series cannot be established from these plots alone.

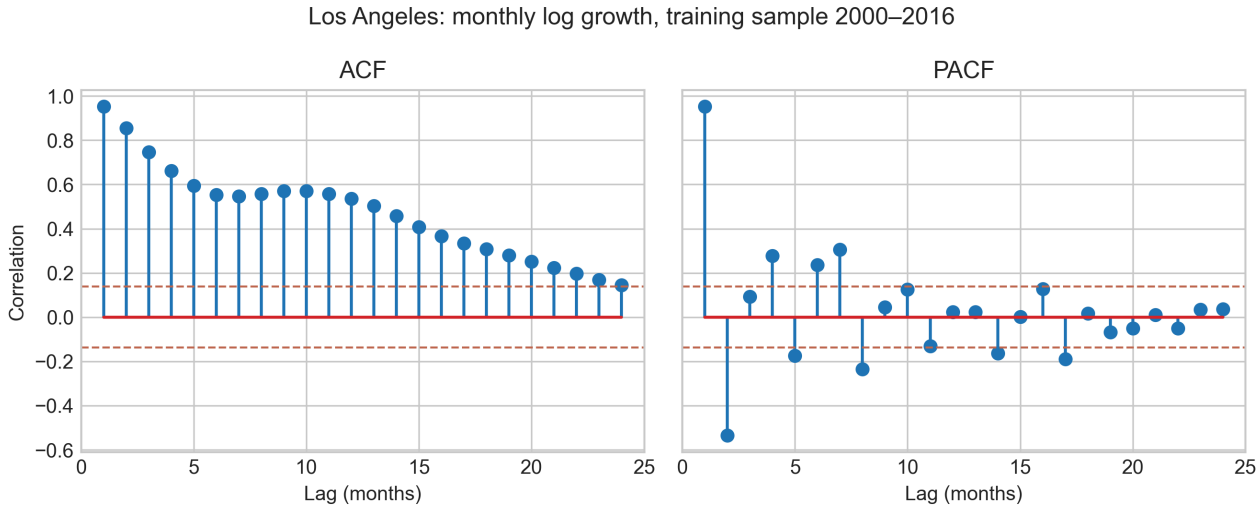


Figure 2. ACF and PACF of Los Angeles City monthly log growth, 2000–2016

The Ljung–Box test at twelve lags [21] checks whether autocorrelation remains in the ARIMA residuals. It rejects the white noise null for 62 of the 87 fits that converged during training. Of the three series in Table 4, Los Angeles City is the only one for which the null is not rejected ($p = 0.133$). The test excludes the first twelve residuals and uses the software's default degrees of freedom without adjustment for p+q.

During the training search, 117 of the 783 candidate fits failed or did not converge. Across 10,092 forecast origins, a converged fallback order was needed 1,589 times (15.7%), at least once in every one of the 87 series. The audit log records the order used at each origin and shows that a city may use different ARIMA orders at successive origins despite the fixed fallback ranking.

The widespread residual dependence prevents a claim of adequate fit to the training data. Model performance is judged mainly by errors outside the estimation sample.

DOI: 10.54254/2754-1169/2026.AB39208

Table 4. Selected ARIMA orders and training diagnostics

Unit	ARIMA order	AIC	BIC	Ljung–Box p(12)
California	(1, 1, 2)	-2082.7	-2066.1	0.003
LA County	(2, 1, 2)	-1818.0	-1798.1	<0.001
LA City	(1, 1, 2)	-1835.8	-1819.2	0.133

4. Pre-test evaluation and test results

The 2017–2019 pre-test results played no part in model selection and are reported on their own, before the 2020–2026 final test.

4.1. Pre-test evaluation, 2017–2019

In the pre-test period, ARIMA had the lowest city-median MASE at one and three months and drift at six and twelve months (Table 5). The longer-horizon result is less reliable. At twelve months only 25 target months can be evaluated. The pre-test errors did not alter the ARIMA ranking, which had been fixed on data through 2016, or the model list and combination weight. Tables 5 and 6 use different MASE denominators and should not be read as a before-and-after comparison.

Table 5. Median city MASE in the 2017–2019 pre-test evaluation period

Model	1 month	3 months	6 months	12 months
Naïve	1.497	4.172	6.743	11.004
Drift	1.327	3.683	5.740	8.032
Holt	0.614	3.141	8.551	18.037
ETS	0.604	3.008	7.549	13.125
ARIMA	0.352	2.203	6.262	10.327
Combo	0.447	2.490	6.854	11.630

4.2. Results across 85 cities

ARIMA's lead is largest at one month, where its city-median MASE is 0.350 against 1.196 for drift (Table 6, Figure 3). By six months the gap has narrowed to 5.509 versus 5.969. At twelve months it reverses: ARIMA's 9.656 is above drift's 8.770. Within-city comparisons show the same pattern. Relative to drift, ARIMA reduces MAE by a median of 71.1% across cities at one month, 40.2% at three, and 7.7% at six. At twelve months its MAE is a median 9.8% higher. These figures are medians of city-level percentages. They will not match ratios of the Table 6 medians, and they carry no significance test.

Table 6. Median city MASE in the 2020–2026 test period

Model	1 month	3 months	6 months	12 months
Naïve	1.311	3.744	6.877	10.908
Drift	1.196	3.430	5.969	8.770
Holt	0.551	2.752	7.455	15.943

DOI: 10.54254/2754-1169/2026.AB39208

Table 6. (continued)

ETS	0.538	2.586	6.587	12.140
ARIMA	0.350	2.020	5.509	9.656
Combo	0.414	2.176	5.923	10.512

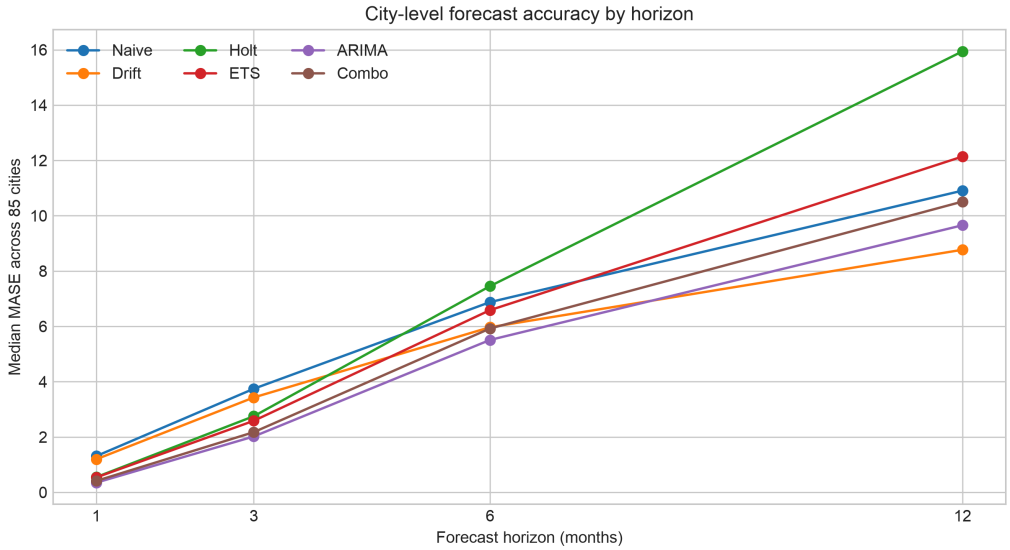


Figure 3. Median city MASE at the four evaluated horizons (1, 3, 6, and 12 months)

4.3. Los Angeles City and regional comparison

Los Angeles City follows the same horizon pattern in dollar terms (Table 7, which also gives percentage errors). ARIMA's one-month MAE was $1,594, well under drift's $6,072. By six months the two were close, at $28,051 and $30,496, and at twelve months drift was ahead, with $43,582 against $48,623 for ARIMA. Figure 4 shows these twelve-month forecasts against their target dates.

Table 7. Los Angeles City forecast errors in the test period

Horizon	Model	RMSE ($)	MAE ($)	MAPE (%)	MASE
1	Naïve	7,785	6,307	0.72	1.270
1	Drift	7,331	6,072	0.69	1.222
1	Holt	3,444	2,808	0.32	0.565
1	ETS	3,364	2,748	0.32	0.553
1	ARIMA	2,156	1,594	0.18	0.321
1	Combo	2,543	1,982	0.23	0.399
3	Naïve	22,159	18,183	2.08	3.660
3	Drift	20,691	17,433	1.98	3.509
3	Holt	17,304	14,093	1.61	2.837
3	ETS	16,289	13,159	1.50	2.649
3	ARIMA	12,727	9,873	1.13	1.988
3	Combo	13,896	10,781	1.24	2.170

DOI: 10.54254/2754-1169/2026.AB39208

Table 7. (continued)

6	Naïve	39,178	32,369	3.70	6.516
6	Drift	35,509	30,496	3.46	6.139
6	Holt	45,391	37,515	4.25	7.552
6	ETS	40,274	33,274	3.77	6.698
6	ARIMA	33,945	28,051	3.17	5.647
6	Combo	36,104	30,048	3.40	6.049
12	Naïve	62,495	49,409	5.68	9.946
12	Drift	52,527	43,582	4.92	8.773
12	Holt	95,986	81,606	9.26	16.427
12	ETS	73,776	60,566	6.90	12.192
12	ARIMA	60,859	48,623	5.54	9.788
12	Combo	64,770	52,219	5.96	10.512

Note. Every model–horizon row covers the same 80 test target months. MAPE is the mean absolute percentage error.

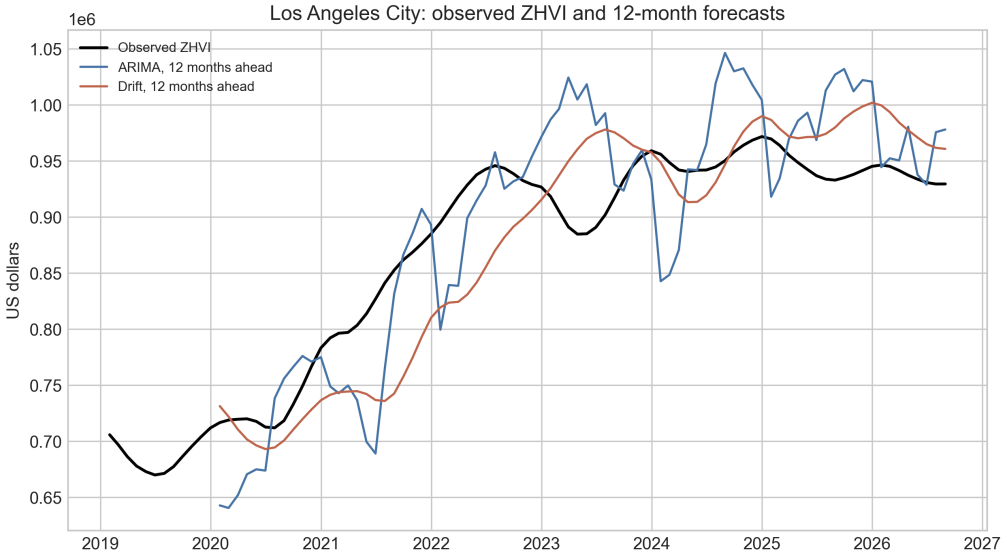


Figure 4. Los Angeles City ZHVI and target-aligned 12-month-ahead forecasts

Note. Each colored point is a separate forecast, made at the origin twelve months before the target date where it is plotted. The lines join forecasts from different origins and do not trace a single forecast path.

Drift also has a lower twelve-month MASE than ARIMA for the state, the county, and Los Angeles City (Table 8), in line with the city medians in Table 6. Because these areas overlap, the agreement is one descriptive check rather than three independent replications, and individual cities may still depart from it.

Table 8. MASE for selected models at state, county, and city levels

Unit	Horizon	Naïve	Drift	ARIMA	Combo
California	1	1.355	1.242	0.269	0.343
California	3	3.966	3.594	1.812	1.968

DOI: 10.54254/2754-1169/2026.AB39208

Table 8. (continued)

California	6	7.420	6.556	5.130	5.337
California	12	12.497	10.666	11.038	11.935
LA County	1	1.258	1.173	0.312	0.393
LA County	3	3.631	3.376	1.873	2.055
LA County	6	6.542	5.946	5.341	5.738
LA County	12	10.382	8.547	9.065	10.159
LA City	1	1.270	1.222	0.321	0.399
LA City	3	3.660	3.509	1.988	2.170
LA City	6	6.516	6.139	5.647	6.049
LA City	12	9.946	8.773	9.788	10.512

4.4. Descriptive model rankings and out-of-sample R^2

Among the four main models, ARIMA ranks first in all 85 cities at one month and in 76 cities at twelve months (Table 9). In each of these cities, the MASE of the model ranked second is at least 2% above ARIMA's. That ranking leaves out drift, whose twelve-month MASE is lower than ARIMA's in 75 cities. ARIMA's first place among the main models therefore does not extend to the drift benchmark at twelve months. Squared-error comparisons against the matched naïve forecast are in Table 10. A positive out-of-sample R^2 means the model beat the naïve forecast on these targets under squared-error loss.

Table 9. Test-period rankings among four main models with a 2% margin

Horizon	Holt	ETS	ARIMA	Combo	Practical ties
1	0	0	85	0	0
3	0	0	84	0	1
6	0	0	82	1	2
12	0	0	76	5	4

Note. Only the four main models are ranked; naïve and drift are left out. A practical tie means the runner-up's MASE is less than 2% above the lowest MASE. Section 2.3 explains the limits of this tolerance.

Table 10. Median out-of-sample R^2 relative to the naïve benchmark

Horizon	Drift	Holt	ETS	ARIMA	Combo
1	0.172	0.814	0.821	0.917	0.888
3	0.188	0.437	0.503	0.676	0.624
6	0.246	-0.196	0.065	0.343	0.247
12	0.368	-1.113	-0.227	0.232	0.088

Note. Values are medians of the city-level out-of-sample R^2. A positive value means lower squared forecast error than the naïve benchmark on the same observations, and a negative value means higher. The MASE rankings in Table 9 are computed separately and do not use this statistic.

DOI: 10.54254/2754-1169/2026.AB39208

4.5. Exploratory Diebold–Mariano comparison

After Holm adjustment, ARIMA's scaled absolute loss is significantly lower than drift's at one and three months for the state, the county, and the equal-weighted city mean (Table 11). The six- and twelve-month differences are not significant. Both the 12-month and the 24-month HAC bandwidths give the same conclusions. For the city test, averaging over cities leaves 80 monthly loss differences; the 6,800 city-month values are not independent observations. The tests remain exploratory because the comparison was chosen post hoc, the reference distribution is approximate, and the losses may be nonstationary.

Under the 12-month bandwidth, the equal-weighted city mean of ARIMA-minus-drift scaled loss is −0.898 [−1.176, −0.620] at one month and −1.465 [−2.171, −0.759] at three months. At six and twelve months it is −0.530 [−2.084, 1.025] and 0.898 [−2.628, 4.424]. Brackets give pointwise 95% intervals, and negative values favor ARIMA. The longer-horizon intervals include zero but are too wide to show that the two models are equally accurate. The twelve-month advantage for drift therefore remains descriptive. These intervals use mean monthly losses, whereas Table 6 reports city medians.

Table 11. ARIMA versus drift: exploratory DM tests on scaled absolute loss

Level	Horizon	Loss diff.	DM statistic	Holm p	Lower loss
California	1	-0.973	-9.04	<0.001	ARIMA
California	3	-1.782	-3.71	0.003	ARIMA
California	6	-1.427	-1.30	1.000	ARIMA
California	12	0.373	0.11	1.000	Drift
LA County	1	-0.861	-9.16	<0.001	ARIMA
LA County	3	-1.502	-3.87	0.002	ARIMA
LA County	6	-0.605	-0.80	1.000	ARIMA
LA County	12	0.517	0.27	1.000	Drift
85-city mean	1	-0.898	-10.17	<0.001	ARIMA
85-city mean	3	-1.465	-4.19	<0.001	ARIMA
85-city mean	6	-0.530	-0.76	1.000	ARIMA
85-city mean	12	0.898	0.51	1.000	Drift

Note. Loss diff. is ARIMA's mean scaled absolute loss minus drift's, so negative values favor ARIMA. Holm p-values are adjusted across all 12 tests and capped at 1.000. This comparison was revised after earlier results had been seen.

5. Heterogeneity and period checks

The full-test pattern is checked below within fixed groups of cities and within two parts of the test period.

5.1. Fixed city groups

The horizon pattern holds within every price and volatility tercile, with groups fixed on data through 2019 (Table 12). In each tercile, ARIMA has the lowest group-median MASE at one, three, and six months, and drift has the lowest at twelve months. Since the groups all come from the same 85

DOI: 10.54254/2754-1169/2026.AB39208

cities, this is a descriptive check, with no test of between-group differences. Group medians can also hide variation among individual cities and say nothing about its economic causes.

Table 12. Lowest group-median MASE among all six models within fixed city groups

Grouping	Tercile	1 month	3 months	6 months	12 months
Dec. 2019 price	Low	ARIMA (0.31)	ARIMA (1.86)	ARIMA (5.18)	Drift (8.20)
Dec. 2019 price	Middle	ARIMA (0.34)	ARIMA (1.97)	ARIMA (5.44)	Drift (8.28)
Dec. 2019 price	High	ARIMA (0.44)	ARIMA (2.55)	ARIMA (6.63)	Drift (9.98)
Historical volatility	Low	ARIMA (0.36)	ARIMA (2.22)	ARIMA (5.78)	Drift (9.47)
Historical volatility	Middle	ARIMA (0.33)	ARIMA (1.96)	ARIMA (5.45)	Drift (8.77)
Historical volatility	High	ARIMA (0.34)	ARIMA (1.91)	ARIMA (5.41)	Drift (8.19)

Note. Terciles are formed with pandas qcut. Price cutoffs (December 2019 ZHVI) are $566,752 and $774,174; growth-volatility cutoffs (standard deviation of monthly log growth, 2000–2019) are 0.010900 and 0.012384. From low to high, the price groups contain 29, 28, and 28 cities and the volatility groups 29, 27, and 29. All 85 volatility values are distinct. The uneven split comes from floating-point rounding: the upper cutoff falls a negligible distance below South Pasadena's value, which places that city in the high group. The supplementary workbook lists each city's group and the full-precision cutoffs.

5.2. Pandemic and post-2022 periods

At one month ARIMA has the lowest pooled city MASE in both subperiods (Table 13). The twelve-month ranking flips. For targets from March 2020 through December 2021, ARIMA's MASE is 8.280 and drift's 10.276. From January 2022 through August 2026, drift is lower, at 8.545 against 10.425. The pandemic period is shorter, and some of its twelve-month forecasts were made before the pandemic started. January and February 2020 are part of the full test but fall in neither subperiod.

Table 13. Pooled mean scaled city errors by test subperiod

Period	Horizon	Naïve	Drift	Holt	ETS	ARIMA	Combo
2020–21	1	1.737	1.486	0.608	0.608	0.385	0.471
2020–21	3	4.816	4.082	3.053	2.854	2.258	2.460
2020–21	6	8.725	7.014	7.948	6.873	5.606	6.144
2020–21	12	14.671	10.276	16.345	12.171	8.280	9.873
2022–26	1	1.244	1.207	0.571	0.553	0.357	0.426
2022–26	3	3.557	3.466	2.859	2.675	2.104	2.283
2022–26	6	6.227	5.949	7.764	6.750	5.848	6.161
2022–26	12	9.631	8.545	17.085	12.471	10.425	10.898

Note. The 2020–21 rows cover target dates from March 2020 to December 2021 (22 months), and the 2022–26 rows cover January 2022 to August 2026 (56 months). January and February 2020 are in the full test but in neither subperiod. Values pool scaled city–month losses, whereas Tables 5 and 6 report medians across cities.

6. Discussion and limitations

ARIMA's advantage at one and three months is the most stable result here: it appears in every price and volatility tercile, in both subperiods, and under both HAC bandwidths. The twelve-month result is weaker. Drift has the lower full-test median and a lower MASE in 75 cities. However, the ranking

DOI: 10.54254/2754-1169/2026.AB39208

flips between subperiods, and the interval for the loss difference includes zero. Its lower error in this sample is therefore no guarantee that it will beat ARIMA in future twelve-month forecasts.

Averaging the ARIMA and ETS forecasts did not reduce the median error across cities compared with ARIMA alone. Bates and Granger [11] showed that combining forecasts can reduce error, but their result does not guarantee an improvement from assigning equal weights to these two models. As in the M3 competition [22], methods are judged by their observed accuracy, and simple benchmarks are retained when assessing performance.

Earlier housing studies documented differences in forecastability across U.S. states [1] and applied dynamic model averaging and selection to all 50 states [2]. Case and Shiller [23] found serial correlation in house price changes. Differences in data, geographic coverage, and evaluation periods prevent a direct numerical comparison with this ZHVI analysis. The finding here is that the relative accuracy of methods varies across horizons within Los Angeles County. ARIMA performs better one to three months ahead, but drift has the lower median error over the full test period at twelve months.

Zillow's methodology notes describe revisions to the entire ZHVI history [8]. Truncating a later vintage at an earlier forecast origin leaves those revisions in the historical values, so it does not recover the information available at that time. The target series is also smoothed and seasonally adjusted. Its dependence over short periods and its forecast errors may therefore differ from those of raw transaction prices. Some of ARIMA's large improvement at one month may reflect these data features, but the analysis does not test whether smoothing or revisions explain the gain.

ARIMA's residual dependence and convergence failures limit conclusions about model adequacy (Section 3.2). The unit root null was again not rejected for any of the 87 differenced series in the repeated ADF tests. A common value of $d = 1$ maintains comparability across series but does not establish stationarity after differencing. Convergence was audited only for ARIMA; Holt and ETS were required only to return finite forecasts.

Differences between the procedures also affect how their accuracy should be interpreted. The transformation back to levels applies the variance estimated for forecasts one month ahead at every horizon and targets the mean. MAE and MASE favor forecasts of the median. The main models use log values, whereas the benchmarks use dollar values. The results thus compare complete forecasting procedures, and differences in accuracy cannot be attributed to model structure alone.

The 85 cities matched to the county's official list using Zillow's state and county identifiers all have complete data over the sample period. They may not represent every housing market in the county. The index also changes construction in 2012, when the older historical index gives way to the series based on neural Zestimates [8]. The training sample spans both constructions, and no test for a structural break in 2012 is reported. The stationarity and residual diagnostics in Section 3 may partly reflect this transition.

Each city receives equal weight regardless of its housing stock or population, and the overlapping state, county, and city series share evaluation dates. Their errors are not independent. The analysis is limited to retrospective point forecasts of the index and does not assess forecasts of transaction prices, investment returns, or the calibration of prediction intervals.

Archiving successive Zillow releases would allow the same comparison to use only information available at each forecast origin. Further analyses could be prespecified for that evaluation, covering transformations back to levels at each horizon as well as prediction intervals and additional predictors.

DOI: 10.54254/2754-1169/2026.AB39208

7. Conclusion

Forecast horizon matters when choosing a method for cities in Los Angeles County. ARIMA had the lowest median MASE through six months, though its advantage over drift was small by six months. At twelve months, drift had the lower median MASE over the full test period. For buyers, sellers, renters, and planners needing estimates one to three months ahead, ARIMA is a reasonable default. Near a year ahead, more complex models should first be compared with naïve and drift benchmarks at the same horizon. These findings rely on a single revised and smoothed ZHVI vintage and on ARIMA fits that sometimes showed residual correlation or failed to converge. Repeating the comparison with archived Zillow releases would test whether the pattern holds using only the information available at each origin.

Data and reproducibility note

The retained materials include Zillow's city, county, and state CSV snapshots dated September 29, 2026, the script run_forecast_analysis.py, forecast records for each origin and target date, and the supplementary workbook. All reported errors can be recalculated from the saved records; software versions and functions are listed in Table A1. Additional HAC checks are recorded in review_checks_v9.py, while the workbook contains the original estimates for Tables 1–13 but excludes the additional sensitivity intervals. The code and forecast records are stored locally, and no public repository is provided.

References

[1] Rapach, D. E., & Strauss, J. K. (2009). Differences in housing price forecastability across US states. *International Journal of Forecasting, 25*(2), 351–372. https://doi.org/10.1016/j.ijforecast.2009.01.009

[2] Bork, L., & Møller, S. V. (2015). Forecasting house prices in the 50 states using dynamic model averaging and dynamic model selection. *International Journal of Forecasting, 31*(1), 63–78. https://doi.org/10.1016/j.ijforecast.2014.05.005

[3] Xu, X., & Zhang, Y. (2021). House price forecasting with neural networks. *Intelligent Systems with Applications, 12*, 200052. https://doi.org/10.1016/j.iswa.2021.200052

[4] Jin, B., & Xu, X. (2025). Forecasts of residential real estate price indices for ten major Chinese cities through Gaussian process regressions. *International Journal of Empirical Economics, 4*(1), 2450013. https://doi.org/10.1142/S2810943024500136

[5] Plakandaras, V., Pragidis, I., & Karypidis, P. (2024). Deciphering the U.S. metropolitan house price dynamics. *Real Estate Economics, 52*(2), 434–485. https://doi.org/10.1111/1540-6229.12453

[6] Zillow Research. (n.d.). Housing data: Zillow Home Value Index [Data set]. Retrieved September 29, 2026, from https://www.zillow.com/research/data/

[7] Los Angeles County Auditor-Controller. (n.d.). Incorporated areas: 88 cities served by the county (fiscal year 2025). Retrieved September 29, 2026, from https://auditor.lacounty.gov/incorporated-areas-cities/

[8] Olsen, S. (2023, February 10). Zillow Home Value Index methodology, 2023 revision: What's changed? *Zillow Research*. https://www.zillow.com/research/methodology-neural-zhvi-32128/

[9] Gardner, E. S., Jr. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting, 4*(1), 1–28. https://doi.org/10.1002/for.3980040103

[10] Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting, 18*(3), 439–454. https://doi.org/10.1016/S0169-2070(01)00110-8

[11] Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly, 20*(4), 451–468. https://doi.org/10.1057/jors.1969.103

[12] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

DOI: 10.54254/2754-1169/2026.AB39208

[13] Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599
[14] Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica, 55*(3), 703–708. https://doi.org/10.2307/1913610
[15] Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting, 13*(2), 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4
[16] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. https://doi.org/10.2307/4615733
[17] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366a), 427–431. https://doi.org/10.1080/01621459.1979.10482531
[18] Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika, 71*(3), 599–607. https://doi.org/10.1093/biomet/71.3.599
[19] Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics, 54*(1–3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y
[20] Hobijn, B., Franses, P. H., & Ooms, M. (2004). Generalizations of the KPSS-test for stationarity. *Statistica Neerlandica, 58*(4), 483–502. https://doi.org/10.1111/j.1467-9574.2004.00272.x
[21] Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*(2), 297–303. https://doi.org/10.1093/biomet/65.2.297
[22] Makridakis, S., & Hibon, M. (2000). The M3-Competition: Results, conclusions and implications. *International Journal of Forecasting, 16*(4), 451–476. https://doi.org/10.1016/S0169-2070(00)00057-1
[23] Case, K. E., & Shiller, R. J. (1989). The efficiency of the market for single-family homes. *American Economic Review, 79*(1), 125–137. https://www.jstor.org/stable/1804778

DOI: 10.54254/2754-1169/2026.AB39208

Appendix A. Software and functions

Table A1. Software versions and functions

Component	Software or function	Version or role
Runtime	Python	3.9.6
Packages	NumPy; pandas; SciPy; statsmodels	2.0.2; 2.3.3; 1.13.1; 0.14.6, respectively
ARIMA	statsmodels.tsa.arima.model.ARIMA	Model fitting
Holt	statsmodels.tsa.holtwinters.Holt	Model fitting
ETS	statsmodels.tsa.holtwinters.ExponentialSmoothing	Model fitting
Tests and diagnostics	adfuller, kpss, acf, pacf	statsmodels.tsa.stattools
Residual test	acorr_ljungbox	statsmodels.stats.diagnostic