

Comparative Efficacy of Holt-Winters, ETS and Croston Methods in Multi-Industry Rolling-Origin Backtesting

Christopher I. V. Farmer, LL.M. (University of Worcester) · Independent Researcher
· civfarmer@gmail.com

Preprint; not yet peer reviewed. Version 1.1, October 2026.

DOI: <https://doi.org/10.5281/zenodo.23462158>

Licence: Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).
Commercial use needs a licence from the author: civfarmer@gmail.com.

Abstract

Operational demand forecasting rarely fails because a single algorithm is weak; it fails because one algorithm is imposed on heterogeneous series whose statistical shapes differ. We study a dependency-free forecasting engine that implements the classical exponential-smoothing family (simple exponential smoothing, Holt’s linear trend, and additive Holt-Winters — the point-forecast equivalents of the additive error-trend-seasonal, or ETS, models), ordinary-least-squares regression with trend, seasonal dummies and exogenous drivers, and Croston’s method for intermittent demand, together with seasonal-naive and moving-average baselines. Rather than committing to one model, the engine performs per-series **rolling-origin (walk-forward) backtesting** and selects a champion by lowest held-out error — **Weighted Absolute Percentage Error (WAPE)** where it is defined, mean absolute error where it is not — retaining challengers for transparency. We apply the engine across five commercial domains with deliberately different operational cadences: brewery, direct-to-consumer (DTC) e-commerce and discrete manufacturing on **weekly** data with a 13-week (one-quarter) operational horizon; software-as-a-service (SaaS) recurring revenue and education enrolment on **monthly** data with a 12-month horizon. Running the engine as published on 30 September 2026 over the 31 benchmark series produces a pattern that is interpretable at the three-step selection horizon: additive Holt-Winters wins where annual seasonality is strong and at least two seasonal cycles of history exist (8 of 31 series); Holt’s linear trend wins for compounding recurring revenue where no genuine annual season exists (5 series: a distributed spirit and four of the five SaaS segments; the fifth chooses simple exponential smoothing); regression captures clean trending series, including structurally declining lines (5 series); robust local-level models (moving average, simple exponential smoothing) win 8 series, most of them lines whose hold-out window falls in a flat stretch of their seasonal cycle; and seasonal-naive wins 4. Croston’s method wins one of the three lines flagged intermittent, for reasons the backtest exposes: on a Q4 gift bundle, whose zeros are almost all scheduled, seasonal-naive is best, or tied for best, over every window tested; on a limited seasonal release, whose in-season arrivals are random, Croston has the lowest absolute error on the hold-out; and on a spares line generated as random batches over a non-zero baseline, Croston reduces to simple exponential smoothing. The pattern is also fragile: every series is judged on just six periods, 17 of the 31 champions hold only a narrow lead by the engine’s own test, and changing only the backtest horizon or the number of origins changes between 13 and 20 of the 31 assignments. We argue that the operational horizon and bucket must be matched to the domain’s decision cadence and lead-time structure, not chosen for statistical convenience; that backtest-driven per-series selection is the honest way to let each series declare its own model, provided its margins and sensitivity are reported with it; and that the selection horizon should match the operational one. All reported figures are computed by executing the reference engine on semi-synthetic benchmark datasets whose seasonal shapes are anchored to public sources and whose

operational detail is clearly-labelled simulation; they should therefore be read as demonstrating *selection behaviour under known generating processes*, not as measured accuracy on proprietary operational data.

Keywords: demand forecasting; exponential smoothing; Holt-Winters; ETS; Croston’s method; intermittent demand; rolling-origin evaluation; walk-forward backtesting; WAPE; model selection; inventory management; recurring revenue

1. Introduction

Applied forecasting practice contains a recurring category error. A team adopts one model family — often whatever a library defaults to, or whatever the most recent competition rewarded — and applies it uniformly to a portfolio of series that in reality have nothing statistically in common. A brewery’s flagship lager, a fast-growing skincare hero SKU, a compounding enterprise SaaS segment, and a trickle of after-sales spare parts are not four instances of the same forecasting problem. They differ in the strength and existence of seasonality, in whether growth is compounding or linear, in the fraction of periods with zero demand, in the length of available history, and — crucially — in the *decision* the forecast must support and the *cadence* at which that decision is taken.

This paper takes the opposite stance: **the model should be a property of the series, discovered empirically, not a property of the vendor, imposed by default**. We examine a compact, dependency-free forecasting engine built around this principle. The engine implements a bank of transparent, classical methods — the exponential-smoothing family, OLS regression with calendar and exogenous features, and Croston’s method for intermittent demand, plus honest baselines — and, for each series, decides among them by rolling-origin backtesting. The champion is whichever model forecasts held-out actuals most accurately, measured primarily by WAPE. Nothing is selected on the basis of in-sample fit, and nothing is selected by name or reputation.

We make four contributions. First, **a reproducible, inspectable engine**: because it has no external dependencies and computes real statistics, every claim here can be reproduced by executing the same code on the same data, both of which are published (§4.3). Second, **a genuinely multi-industry evaluation** across five domains configured independently, each with its own seasonality assumption, metrics, constraints and — deliberately — its own granularity and horizon. Third, **an argument that horizon must match domain**: weekly buckets with a 13-week horizon are correct for brewery, DTC and manufacturing because supplier lead times are quoted in weeks, whereas monthly buckets with a 12-month horizon are correct for SaaS and education because recurring revenue and termly enrolment are month-native. Fourth, **an account of which method wins where, and why — and of how far that account can be trusted**: champion selection is connected to seasonality strength, intermittency structure and the position of the hold-out window in the seasonal cycle; the presence of zeros proves not to be a reason in itself for Croston’s method, which wins where the hold-out falls among random in-season arrivals, loses where it falls among scheduled zeros and reduces to exponential smoothing where demand never falls to zero; and the champion assignments, though exactly reproducible, prove sensitive to the backtest’s horizon and number of origins.

The remainder of the paper reviews related work (§2), specifies the engine and evaluation protocol (§3), describes the five domains and the experimental design (§4), reports results computed by running the engine, with their sensitivity to the backtest configuration (§5), discusses the drivers of model choice and the horizon-matching argument (§6), and states limitations (§7) before concluding (§8).

2. Related work

Exponential smoothing and the ETS family. The methods at the core of this study descend from the exponentially weighted moving average and its extensions to trend and seasonality by Holt (1957, reprinted 2004) and Winters (1960). For decades these recursions were treated as ad hoc; Ord, Koehler and Snyder (1997) and Hyndman, Koehler, Snyder and Grose (2002) placed them on a rigorous statistical footing by showing that each smoothing method corresponds to a state-space model with a single source of error, giving the now-standard error-trend-seasonal (ETS) taxonomy and an automatic selection procedure based on information criteria. Gardner (2006) surveys the state of the art. Our engine implements the *point-forecast recursions* of the additive ETS sub-family — ETS(A,N,N) as simple exponential smoothing, ETS(A,A,N) as Holt’s linear trend, and ETS(A,A,A) as additive Holt-Winters — fitting smoothing parameters by grid-search minimisation of one-step in-sample squared error. It does not perform full likelihood-based ETS estimation with automatic error/trend/season-type selection; instead it competes the smoothers against other families and selects among them by out-of-sample backtest, a philosophy analogous to automatic ETS but driven by held-out error rather than in-sample penalised likelihood.

Intermittent demand. When demand occurs sporadically, with many zero periods, ordinary smoothing of the raw series is biased because it decays the estimate during runs of zeros. Croston (1972) proposed decoupling the problem into two separately smoothed processes — the size of non-zero demand and the interval between demands — and combining them into a demand-rate estimate. Syntetos and Boylan (2001) showed that Croston’s estimator is itself biased, and Syntetos and Boylan (2005) tested the bias-corrected variant now widely used as the Syntetos-Boylan Approximation (SBA), which outperformed Croston’s method on 3,000 intermittent demand series from the automotive industry; Syntetos, Boylan and Croston (2005) gave an operational classification of demand patterns (by average inter-demand interval and the squared coefficient of variation of demand sizes) that tells a practitioner *when* intermittent methods should be preferred. Our engine implements classical Croston and always competes it on series flagged or detected as intermittent; a production port would add the SBA bias correction. A central empirical finding below is that the presence of zeros is not sufficient justification for Croston: the *structure* of the intermittency matters.

Rolling-origin evaluation. The methodological backbone of this work is out-of-sample evaluation. Tashman (2000) reviews the design of out-of-sample accuracy tests and argues that rolling-origin evaluation — repeatedly advancing the forecast origin through the series and re-forecasting — yields more efficient and reliable accuracy estimates than a single fixed split. Bergmeir and Benítez (2012) compare cross-validation with evaluation on the last part of each series and find that cross-validation gives the more robust model selection, recommending a blocked form of it as standard practice for time series. We adopt an expanding-window rolling-origin protocol as the sole basis for champion selection; as the deployed dashboards configure it, the protocol scores only the last six periods of each series, close to the last-block evaluation those authors found less robust — a point §5.4 bears out.

Accuracy metrics. Hyndman and Koehler (2006) showed that many popular accuracy measures degenerate in common situations — percentage errors explode near zero actuals, and symmetric variants have their own pathologies — and recommended scale-free measures such as the Mean Absolute Scaled Error (MASE). For portfolios reported to operators in units of demand and cash, we report WAPE (the sum of absolute errors divided by the sum of absolute actuals, also called weighted MAPE), which is robust to the scale of individual series and to occasional zeros, and rank on mean absolute error, which orders models identically whenever WAPE is defined (§3.2), while retaining sMAPE and MASE-adjacent diagnostics. WAPE and closely related weighted errors are the reporting standard in large-scale retail forecasting; the M5 competition (Makridakis, Spiliotis and Assimakopoulos, 2022), run on 42,840 hierarchical Walmart unit-sales series, forecast at the 30,490 bottom-level series and summed to the levels above, evaluated entries on a weighted, scaled error. The comprehensive treatment of

all of these methods in Hyndman and Athanasopoulos (2021) is the standard reference and informs the engine’s design.

Positioning. Relative to this literature, the present work is not a new algorithm but an *engineering-and-evaluation* study: it asks what happens when a disciplined per-series backtest-and-select procedure, built from well-understood classical components, is applied honestly across five operationally distinct domains, and reports the resulting champion assignments, their explanations and their stability.

3. Methods

3.1 The forecasting engine

The engine is a single dependency-free module; the version used here is the one published on the author’s site since 30 September 2026 (§4.3). Series are plain arrays of numbers, oldest first; a series carries a seasonal period s (12 for monthly, 52 for weekly) supplied by the domain configuration. All models expose a common interface — a name, fitted parameters, in-sample fitted values (used to estimate residual dispersion), and a `forecast(h)` function — so they can be competed uniformly.

Model families.

- **Seasonal-naïve.** Forecast for period $t+h$ equals the observed value one full season earlier. An honesty baseline; strong when a series simply repeats a fixed seasonal pattern.
- **Moving average.** Forecast equals the mean of the last w observations (default $w=3$). A robust local-level estimator for noisy series with weak structure.
- **Simple exponential smoothing (SES).** Level-only smoother; smoothing parameter α chosen by grid search over 0.05–0.95 minimising in-sample sum of squared one-step errors (SSE). Equivalent to the ETS(A,N,N) point forecast.
- **Holt’s linear trend.** Level and additive trend, parameters (α, β) selected by two-dimensional grid search on SSE. Equivalent to ETS(A,A,N). Suited to compounding or steadily declining series.
- **Additive Holt-Winters.** Level, additive trend and additive seasonal components with parameters (α, β, γ) selected by three-dimensional grid search on SSE; requires at least two full seasonal cycles of history. Equivalent to ETS(A,A,A).
- **OLS regression (trend + season + drivers).** A linear model with intercept, a linear time trend, $s-1$ seasonal dummy variables, and optional exogenous drivers (temperature, promotions, ad-spend, PMI, marketing spend, pipeline). Coefficients are estimated by ridge-regularised normal equations (penalty 10^{-4} , intercept unpenalised) for numerical stability. At weekly resolution the design matrix carries 53 calendar columns before drivers. The backtest treats the drivers differently from deployment (§3.2).
- **Croston’s method (intermittent demand).** Separate exponential smoothing ($\alpha=0.1$) of non-zero demand sizes and inter-arrival intervals, combined into a constant demand-rate forecast; as in Croston’s original, α is fixed and both smoothers start from the first demand — its size, and the number of periods up to and including it (Hyndman and Athanasopoulos, 2021, §13.2). Classical Croston is implemented; the SBA bias correction is noted as the production upgrade.

Metrics. For aligned held-out actual/prediction pairs the engine computes MAE, RMSE, MAPE, sMAPE, **WAPE** (headline), bias, and percentage bias. WAPE is defined as $\sum |\text{prediction}_i - \text{actual}_i| / \sum |\text{actual}_i| \times 100$. When the held-out actuals sum to zero over a backtest window — possible for heavily intermittent series — WAPE is undefined, and the engine ranks and reports on MAE (§3.2).

Prediction intervals. From the champion’s in-sample residual standard deviation σ , the

engine forms 80% and 95% bands using normal quantiles, widening them with horizon by the fixed factor $\sqrt{1 + 0.4(h - 1)}$ at step h , and flooring lower bounds at zero for non-negative demand. The factor is a heuristic applied to every family alike: it matches the ETS(A,N,N) variance growth $\sigma^2[1 + \alpha^2(h - 1)]$ (Hyndman and Athanasopoulos, 2021, §8.7) only when $\alpha^2 = 0.4$, whatever α was fitted, and for Croston’s method no algebraic interval exists at all (Hyndman and Athanasopoulos, 2021, §13.2).

Recalibration. When a new actual arrives, the engine appends it, classifies it against the prior 95% interval and residual scale as *normal*, *anomalous* (beyond $\sim 1.5\sigma$) or *structural* (outside the interval), re-estimates the level and forecast with the incumbent champion, and re-runs selection, reporting whether the incumbent is still best and which model would replace it; it does not switch champions by itself. Recalibration is part of the operational system but is not exercised in the static evaluation reported here.

3.2 Rolling-origin backtesting and champion selection

Model selection is performed exclusively out-of-sample by expanding-window rolling-origin evaluation. Given a series of length n , a backtest horizon H , and a fold count F , the engine defines fold origins at cut points $n - H - i$ for $i = 0, \dots, F-1$ (subject to a minimum training length of $\max(H+1, \lfloor n/2 \rfloor)$), and, at each origin, fits the model on all observations up to the cut and forecasts H steps ahead; a model that cannot be fitted at every origin (Holt-Winters with fewer than two seasonal cycles of training data, for instance) is dropped from the comparison. The held-out predictions and actuals are pooled across folds and scored once. In the configuration used by the operational dashboards and reproduced here, $H = 3$ and $F = 4$, giving four contiguous, recent, expanding-window origins each evaluated three steps ahead — twelve held-out points per series. Because the origins are one period apart, those twelve points cover only six distinct periods, the middle ones counted two or three times: the weeks beginning 12 May to 16 June 2025 in the weekly domains, and January to June 2025 in the monthly ones. Every series in a domain is judged on the same six periods.

Candidate models are restricted to the domain’s **model policy** (a declared list of eligible families; see §4) — in this corpus the same six families in every domain — with Croston additionally forced onto any series that is flagged intermittent or whose zero fraction exceeds 30%. Each eligible candidate is backtested, and candidates are ranked by mean absolute error (MAE), with root mean squared error breaking ties. Because WAPE is the MAE divided by the mean absolute actual of the same held-out set, the order is the same as WAPE’s whenever WAPE is defined, and WAPE is the score reported; when the held-out actuals are all zero, the MAE itself is reported. The lowest-error model becomes the champion and the runner-up is retained as a challenger with a human-readable selection reason, which labels a lead of less than 1.2 WAPE points (on MAE, no more than the larger of 0.5 units and 5% of the champion’s error) a narrow one. This makes model choice a transparent, auditable consequence of measured held-out error rather than an analyst’s prior.

One property of the backtest must be stated plainly. The selection routine passes the driver-augmented regression no future driver values, so at every held-out step the regression is fed each driver’s mean over the whole history (the held-out periods included) instead of the value the driver actually took; at deployment the same model receives the planned driver values. The backtest therefore scores the regression under different inputs from those it is used with, and its score says little about the deployed model (§5.3, §6.5, §7).

A deliberate design property deserves emphasis: **the selection horizon ($H = 3$) is shorter than the operational horizon** at which forecasts are actually issued (13 weeks or 12 months). Selection is therefore performed at a short horizon and the champion is then extrapolated to the operational horizon. We return to the consequences of this choice in §5.4, §6 and §7.

4. Experimental design

4.1 Five domains, five cadences

The engine is instantiated across five domains. Each was configured independently with its own company, metrics, constraints, seasonal assumptions, model-policy rationale, and — the design lever central to this paper — its own granularity and horizon. Table 1 summarises the configurations.

Table 1. Domain configurations.

Domain (company)	Mode	Granularity	Season s	Op. hori- zon	History	Entities	Model-policy headline	Public reference sources
Brewery (Hopmarket Brewing Co.)	inventory	weekly	52	13 wk	156 wk (3 yr)	8 SKUs	Holt-Winters · ETS · weather- regression · Croston	FRED MRTSSM4453USN; Boston Beer (NYSE: SAM); BBPA
DTC e-commerce (Kindred Goods)	inventory	weekly	52	13 wk	156 wk	6 SKUs	Holt-Winters · ETS · promo- regression · Croston	FRED ECOMNSA (quarterly: Q4 peak only); Warby Parker (WRBY); FIGS
Manufacturing (Nexus Devices)	inventory	weekly	52	13 wk	156 wk	6 lines	Holt-Winters · regression · Croston · capacity- constrained	Sonos (SONO) and GoPro (GPRO) filings (Q4 seasonality); FRED DGORDER and ISM PMI (seasonally adjusted: demand cycle and PMI driver only)
SaaS (Cadence)	subscription	monthly	12	12 mo	30 mo	5 seg- ments	Cohort retention · Holt trend · regression · conservative baselines	monday.com (MNDY); HubSpot (HUBS); KeyBanc/OpenView (churn and retention benchmarks)
Education (Northwind Academy)	capacity	monthly	12	12 mo	30 mo	6 courses	Holt-Winters · regression on leads/marketing · capacity- constrained	HESA; UK DfE; Office for Students

The policy headlines describe how each domain’s forecasts are explained and then constrained; the eligible model families are the same in all five (§3.2). The three inventory domains run on **weekly** buckets with a **13-week** (one-quarter) operational horizon, because breweries, online retailers and hardware manufacturers place replenishment or component orders on a weekly cycle against supplier lead times quoted in weeks (typically 4–8). A monthly bucket cannot time a reorder against a five-week lead time; the weekly quarter horizon is the shortest window that spans the longest lead time plus a review period. The two non-inventory domains run on **monthly** buckets with a **12-month** horizon, because their governing quantities — MRR, ARR, churn, net revenue retention and cash runway for SaaS; termly intakes and tutor capacity for education — are month-native, and a weekly bucket has no operational meaning (MRR is billed monthly; tutors are not staffed per week). Section 6 develops this horizon-matching argument.

4.2 Data provenance and the honesty of the benchmark

The five datasets are **semi-synthetic** and are labelled as such throughout the source system. Their *seasonal shapes and scales* are anchored to real public data where such data exist — for example, alcohol-retail seasonality from the U.S. Census/FRED series MRTSSM4453USN and Boston Beer public filings for the brewery; the fourth-quarter e-commerce peak from FRED ECOMNSA (a quarterly series, so the monthly profile within the year is an assumption) and public DTC filings; Q4 hardware seasonality cross-checked against Sonos and GoPro filings for manufacturing, whose durable-goods (FRED DGORDER) and PMI references, both seasonally adjusted, inform only the demand cycle and the PMI driver; public SaaS churn/retention benchmarks for the SaaS segments; and HESA intake seasonality for education. The *operational detail* layered onto those shapes — per-SKU weekly volumes, inventory, lead times, per-segment MRR, per-course enrolment — is deterministically generated simulation, seeded for reproducibility, and is explicitly flagged as simulated wherever it appears. Documented, forecast-changing events are overlaid (a July-2023 heatwave, an August-2024 city festival, a pub-group contract from April 2025 for the brewery; a Black-Friday lift and a flash sale for DTC; a component-shortage output cut and a pre-order ramp for manufacturing; an enterprise-logo step and a self-serve churn quarter for SaaS; a corporate-training framework agreement for education).

The consequence for interpretation is important and we state it plainly. Because the data-generating process is synthetic, the WAPE figures in §5 are **not** empirical accuracy measurements on real proprietary operational series. They are, however, **not fabricated**: every number is produced by executing the actual reference engine on the actual benchmark series under the protocol of §3.2. They therefore faithfully demonstrate how the selection machinery behaves when the underlying structure — seasonality strength, trend, intermittency — is known by construction, which is exactly what a methodology study needs. It also means that the explanations offered in §6 can be checked against the code that generated each series, and they have been.

4.3 Protocol and reproducibility

For every entity in every domain, the engine was invoked exactly as the production dashboards invoke it: `autoSelect` with the domain’s season, the domain’s model-policy family list, exogenous drivers for entities configured to use them, a backtest horizon of 3, and four folds; followed by `buildForecast` at the full operational horizon (13 weeks or 12 months) for the point forecast and prediction interval. The engine is the version published on the author’s site since 30 September 2026, run on the data modules published beside it (see the Reproducibility statement). Champion, champion WAPE (or MAE), the full challenger ranking, the next-period point forecast and its 95% interval were recorded. To test how stable the assignments are, the selection was repeated with twelve origins and with the backtest horizon set to the operational horizon (§5.4), and the driver-augmented regression was re-scored with the held-out periods’ actual driver values (§5.3). The evaluation is deterministic: re-running the engine reproduces every figure below exactly.

5. Results

5.1 Champion assignments and accuracy

Table 2 reports, for all 31 series, the selected champion, its backtested WAPE (three-step-ahead, four rolling origins), its lead over the runner-up, the next-period point forecast and its 95% prediction interval. Two heavily intermittent lines have WAPE undefined on their backtest window (all held-out actuals were zero); for these the engine ranked by MAE, shown

in brackets. The only champion that differs from Version 1.0 of this paper is the Solstice Saison’s: the engine’s earlier sMAPE fallback gave it to seasonal-naive, its MAE rule gives it to Croston (§5.3, §6.3).

Table 2. Per-series champions and backtested error (engine output). Lead = the champion’s margin over the runner-up in WAPE points (in MAE units where WAPE is undefined). Holt = Holt’s linear trend; SES = simple exponential smoothing; Regression = OLS on trend and seasonal dummies, “+ drivers” where exogenous drivers are used.

Domain	Entity	Champion	WAPE %	Runner-up (lead)	Next-period forecast (95% PI)
Brewery	Helios Lager (flagship, summer-peaking)	Holt-Winters	7.5	Holt (+2.66)	566 cases (470–662)
Brewery	Tropic IPA (fast-growing, +30%/yr)	Holt-Winters	3.6	Holt (+2.14)	332 (262–402)
Brewery	Midnight Stout (winter line)	Seasonal-naive	5.6	Regression (+0.87)	83 (53–113)
Brewery	Harvest Pale Ale	Holt-Winters	8.8	Regression (+0.99)	175 (137–213)
Brewery	Solstice Saison (limited release)	Croston	n/a (MAE 14.1 cases)	Seasonal-naive (+3.61 cases)	14 (0–91)
Brewery	Aurora Gin (Christmas gifting)	Holt	7.1	SES (+0.14)	50 (35–64)
Brewery	Orchard Cider (strong summer peak)	Holt-Winters	3.9	Holt (+1.34)	262 (215–309)
Brewery	Clear Skies 0.5% (+38%/yr, Dry-Jan)	Regression	4.5	Holt-Winters (+0.32)	172 (157–187)
DTC	Aurora Serum (hero SKU)	Moving average	7.5	Holt (+1.48)	1,128 units (730–1,525)
DTC	Lumen Candle (new launch, +55%/yr)	Regression	6.4	Holt-Winters (+0.41)	437 (370–504)
DTC	Everyday Tote (bestseller)	Moving average	6.5	SES (+0.59)	764 (464–1,064)
DTC	Trail Bottle (summer accessory)	Seasonal-naive	3.9	Regression (+0.07)	745 (622–868)
DTC	Summit Jacket (winter apparel)	Moving average	7.3	Holt (+0.48)	229 (89–368)
DTC	Holiday Gift Set (seasonal bundle)	Seasonal-naive	n/a (MAE 0)	Moving average (tie)	0 (0–184)
Manufacturing	Nexus Hub (flagship device)	Moving average	5.1	Regression + drivers (+0.07)	506 units (415–598)
Manufacturing	Edge Cam (new launch, +48%/yr)	Holt-Winters	7.6	Moving average (+1.53)	385 (321–449)

Domain	Entity	Champion	WAPE %	Runner-up (lead)	Next-period forecast (95% PI)
Manufacturing	Aura Sensor (volume line)	SES	5.4	Moving average (+1.13)	837 (718-956)
Manufacturing	Nexus Pro (premium, Q4-weighted)	Moving average	6.0	SES (+0.32)	213 (173-254)
Manufacturing	Hub v1 legacy (EOL, -12%/yr)	Regression	4.5	SES (+2.66)	142 (116-167)
Manufacturing	Service Spares (after-sales)	Seasonal-naive	37.6	Regression (+1.22)	70 (0-143)
SaaS	Enterprise (lumpy, low churn)	SES	2.7	Holt (+1.84)	£215,095 (204,312-225,878)
SaaS	Mid-market (compounding core)	Holt	1.2	Regression (+6.00)	£209,143 (207,588-210,697)
SaaS	Self-serve SMB (churny)	Holt	1.5	SES (+8.75)	£188,228 (185,347-191,110)
SaaS	Partners/API (+7.5%/mo)	Holt	2.4	Holt-Winters (+9.29)	£191,570 (190,097-193,044)
SaaS	Legacy plan (declining)	Holt	1.4	SES (+2.56)	£17,412 (17,105-17,719)
Education	Data Analytics Bootcamp (+26%/yr)	Holt-Winters	9.9	Regression + drivers (+6.76)	93 students (67-118)
Education	Cloud Engineering Cert (+30%/yr)	Holt-Winters	8.2	Regression (+0.02)	53 (41-66)
Education	Intro to Coding (high volume)	Holt-Winters	4.0	Regression (+0.46)	104 (84-125)
Education	UX Design Diploma (steady)	Regression	2.3	Holt-Winters (+0.65)	47 (44-50)
Education	Business Foundations (-6%/yr)	Regression	4.1	Holt-Winters (+0.57)	27 (25-29)
Education	Leadership (corporate, contract-led)	SES	27.8	Holt (+0.55)	70 (53-88)

Many of these leads are small. Seventeen of the 31 champions lead by less than the engine’s own “narrow lead” margin of 1.2 WAPE points or tie outright (the gift set); nine lead by less than half a point, and two — the Nexus Hub (5.06 against 5.13) and the Cloud Engineering Cert (8.21 against 8.24) — are dead heats at the precision of the WAPE column.

5.2 Distribution of champions

Across the 31 series the champion distribution is:

Table 3. Champion frequency across all domains.

Champion family	Wins	Where
Additive Holt-Winters	8	Brewery lager, IPA, pale, cider; Manufacturing Edge Cam; Education bootcamp, cloud, intro
Holt (linear trend)	5	Brewery gin; SaaS mid-market, self-serve, partners, legacy
OLS regression (trend+season)	5	Brewery low/no; DTC candle; Manufacturing legacy; Education UX, business
Moving average	5	DTC serum, tote, jacket; Manufacturing hub, pro
Seasonal-naive	4	Brewery stout; DTC bottle, gift set; Manufacturing spares
Simple exponential smoothing	3	Manufacturing sensor; SaaS enterprise; Education leadership
Croston	1	Brewery saison (ranked on MAE); also tested on the gift set and the spares, where it loses

No single family dominates: seven families win at least once, and the largest, Holt-Winters, wins only 8 of the 31 series, so a uniform Holt-Winters deployment would have departed from the backtest’s choice for roughly three-quarters of the portfolio (23 of 31). Croston, forced onto the three intermittent lines, wins one of them; §6.3 explains why the three cases differ. Two cautions apply to every count in Table 3: many of the leads behind it are narrow (Table 2), and the distribution itself depends on the backtest configuration (§5.4).

5.3 The model arena: four illustrative competitions

The value of backtesting is visible in the full challenger rankings. Four examples (WAPE %, best first, unless stated) are instructive.

Helios Lager (brewery flagship, strong annual season, weekly). Holt-Winters **7.5** | Holt 10.1 | SES 10.7 | Moving average 11.2 | Seasonal-naive 21.9 | Regression+drivers 28.7. Holt-Winters wins cleanly by modelling the 52-week seasonal state. The *driver-augmented regression is the worst model here*, but not because of its 53-column seasonal design: in the backtest the engine feeds it each driver’s full-history mean (10.3 °C for temperature) rather than the held-out weeks’ values (14–17 °C in May and June), whereas at deployment it receives the planned values (§3.2). Because the synthetic temperature driver is almost a function of the calendar ($R^2 = 0.96$ on the model’s own trend and seasonal columns), part of the lager’s summer lift sits in the temperature coefficient, and the mean removes it from the forecast. Scored with the held-out weeks’ actual temperatures and promotions, the same regression has a WAPE of 14.4%; without drivers, 13.6%.

Service Spares (manufacturing after-sales, flagged intermittent). Seasonal-naive **37.6** | Regression 38.8 | Holt-Winters 55.2 | Moving average 71.3 | SES 72.4 | **Croston 72.4** | Holt 74.6. Every model struggles, but not for the reason the line’s label suggests. The series has no zero weeks: the generator adds, in any week with probability 0.45, a batch of between 29 and 67 units to a constant base of 15 — random arrivals of random size, the process Croston’s method was built for, but on a non-zero baseline. On a series without zeros Croston’s interval estimate never moves from one period, so the method reduces to simple exponential smoothing with $\alpha = 0.1$, which is also the value the grid search picks for SES here; hence the identical 72.4. Seasonal-naive’s win belongs to these six weeks rather than to the series (§6.3).

Partners/API (SaaS, compounding growth, monthly). Holt **2.4** | Holt-Winters 11.6 | SES 14.0 | Moving average 19.6 | Regression 20.3 | Seasonal-naive 57.7. Holt-Winters was eligible

(30 months $\geq$ two seasonal cycles) and competed, but lost by roughly a factor of five to Holt's trend model — there is no genuine annual season in recurring revenue for Holt-Winters to exploit, so it fits and extrapolates seasonal noise.

Solstice Saison (brewery limited release, flagged intermittent; MAE in cases per week). Croston **14.1** | Seasonal-naive 17.8 | SES 23.0 | Moving average 24.8 | Regression 29.5 | Holt 36.1 | Holt-Winters 41.6. All six held-out weeks happened to see no demand, so WAPE is undefined and the engine ranks on absolute error. Croston's constant 14.1 cases a week is the smallest miss; seasonal-naive forecasts zero for five of the six weeks but repeats last year's 71-case batch for the week of 2 June, which the overlapping folds count three times. Under the earlier sMAPE fallback seasonal-naive won and Croston tied for last, because sMAPE scores any non-zero forecast against a zero actual at its 200% maximum, whatever its size, and seasonal-naive's nine exact zeros decided the ranking.

5.4 Sensitivity of the champion assignments

Table 2's assignments come from one backtest configuration. Table 4 repeats the selection on the same engine and data with more origins, with the backtest horizon set to the operational horizon (13 weeks, 12 months), and with both.

Table 4. Champions under alternative backtest configurations (same engine, same data).

Backtest configuration	Periods scored per series (weekly / monthly)	Champions differing from Table 2 (weekly / monthly)	Champions by family
$H = 3, 4$ origins (Table 2)	6 / 6	—	Holt-Winters 8, Holt 5, regression 5, moving average 5, seasonal-naive 4, SES 3, Croston 1
$H = 3, 12$ origins	14 / 14	13 of 31 (9 / 4)	Regression 12, SES 5, Holt 5, Holt-Winters 3, moving average 3, seasonal-naive 2, Croston 1
$H = 13$ weeks or 12 months, 4 origins	16 / 15	16 of 31 (10 / 6)	Regression 14, Holt-Winters 9, moving average 3, Holt 3, seasonal-naive 2
$H = 13$ weeks or 12 months, 12 origins (4 possible monthly)	24 / 15	20 of 31 (14 / 6)	Regression 19, Holt-Winters 6, Holt 3, moving average 2, seasonal-naive 1

Ten assignments survive all three changes: the Helios Lager, Tropic IPA, Clear Skies, Holiday Gift Set and Hub v1; the mid-market, partners and legacy SaaS segments; and the UX Design and Business Foundations courses. The other 21 change under at least one. Two features of the table need reading with care. In the monthly domains, twelve origins or a twelve-month horizon leave the earliest training windows shorter than the 24 months Holt-Winters needs, and the engine then drops Holt-Winters from the comparison; three of the monthly changes (the bootcamp, cloud and intro courses, all Holt-Winters champions in Table 2) are this exclusion rather than a measured defeat. And regression's gains at the operational horizon come almost entirely from the regression without drivers (12 of its 14 wins with four origins), so they do not depend on the driver treatment described in §3.2. The near-even spread of Table 3 is a property of the three-step, four-origin window; at the operational horizon regression is the most frequent champion.

6. Discussion

6.1 Seasonality strength selects Holt-Winters — when there is enough history

Every one of the eight Holt-Winters champions has genuine annual seasonality *and* at least two full seasonal cycles of history: the brewery’s weather-driven lager, IPA, pale ale and cider (156 weekly observations at $s=52$), the manufacturer’s Q4-weighted new camera, and the academy’s September-intake bootcamp, cloud certificate and intro course (30 monthly observations at $s=12$). Holt-Winters wins because it estimates and projects the seasonal component explicitly; competitors that ignore season (SES, Holt, moving average) miss the seasonal movement across the hold-out window, and seasonal-naïve, which captures the shape, cannot adapt its level or trend and so trails on series that are also growing. The strength of the result varies. The lager and IPA keep Holt-Winters under every configuration in Table 4; the pale ale, whose season is mild (multipliers 0.90–1.14), leads regression by less than a point and loses the title to SES with twelve origins and to regression at the operational horizon with twelve origins; and the three academy courses cannot be re-tested with longer windows at all, because their earliest training windows would be shorter than the 24 months Holt-Winters needs. The two-cycle requirement is not incidental — it is why Holt-Winters is ineligible on very short series, and in this corpus it shapes the monthly results of Table 4.

6.2 Compounding growth without a season selects trend models

The cleanest single result is the SaaS column: **all five recurring-revenue segments choose a non-seasonal model** (Holt for the four growth/decline segments, SES for the enterprise segment), and Holt-Winters — though eligible — loses every time, typically by a large margin. Recurring revenue compounds month over month and has no true annual seasonality; a seasonal model asked to explain it will fit spurious seasonal structure into what is really trend plus noise, and pay for it out of sample. Holt’s linear trend, by contrast, is almost exactly the right generating assumption for a book of MRR growing or declining at a roughly constant rate, which is why its backtested WAPE for the mid-market, self-serve, partners and legacy segments falls between 1.2% and 2.4%, and why the mid-market, partners and legacy segments keep Holt under every configuration in Table 4. The brewery’s Aurora Gin also chooses Holt, but this is a different result: the gin has a strong Christmas peak, and in a hold-out window far from it Holt, SES and the moving average finish within 0.25 points of one another; at the thirteen-week horizon with four origins Holt-Winters wins. The operational lesson is blunt: **do not put a seasonal model on a series that has no season**, and let the backtest enforce it.

6.3 Intermittency structure, not the mere presence of zeros, decides Croston

The three lines flagged intermittent show why the flag settles nothing by itself: each meets Croston’s method in a different way, and the backtest exposes the difference. Croston (1972) models demand that arrives at *random* intervals with *random* sizes — a memoryless arrival process — and estimates a smoothed rate.

The DTC **Holiday Gift Set** is the line whose zeros are almost all *scheduled*: the generator makes it only in October–December, so 117 of its 129 zero weeks fall out of season. The hold-out weeks fall out of season, every model that forecasts zero scores zero, and seasonal-naïve heads a three-way dead heat with the moving average and SES only because it is listed first; Croston loses because it carries the rate it last estimated in December (67.4 a week) into the summer. Over longer windows the result is genuine: seasonal-naïve is best with fifty-two origins (WAPE 52.1% against Croston’s 162.1%) and at a thirteen-week horizon over forty origins (52.1% against 138.2%). For scheduled zeros, seasonal-naïve is close to optimal — it

predicts zero out of season and last year's quantity in season — while Croston's constant-rate estimate smears demand across periods where none occurs.

The brewery's **Solstice Saison** mixes the two kinds of zero. Out of season (January–March and July–October) its zeros are scheduled; in season a batch drops in any week with probability one half, so arrivals are random, and 39 of its 132 zero weeks fall in season. The six hold-out weeks are in season and happened to see no demand, and on absolute error Croston's constant rate is the smallest miss (\$5.3). With twelve origins Croston is still first (WAPE 167.2% against seasonal-naive's 218.0%), but once the evaluation covers a full year, the long July–October off-season included, seasonal-naive wins: 135.1% against Croston's 180.5% with fifty-two origins, and 127.2% against 171.3% at a thirteen-week horizon over forty origins. The selection also has a cost at deployment that the six-week window cannot see. The thirteen weeks the engine then forecasts (beginning 23 June to 15 September 2025) are almost all out of season, and the generator's expected demand over them is about 80 cases, all in the two June weeks; Croston projects 14.1 cases every week, 184 in all, against seasonal-naive's 60.

The manufacturer's **Service Spares** line is the reverse of what its label suggests. It has no zero weeks at all: it is random batches on a non-zero baseline — the memoryless arrival process Croston assumes, without the zeros — and on such a series Croston reduces to simple exponential smoothing (\$5.3). Seasonal-naive's six-week win does not survive a longer window: with twelve origins it ranks fourth of seven, and at a thirteen-week horizon over forty origins Croston ranks first (60.5%, against the moving average's 60.7% and seasonal-naive's 78.4%).

This connects directly to the demand-categorisation framework of Syntetos, Boylan and Croston (2005): the decision to use an intermittent method should follow from measured inter-demand interval and demand-size variability, not from a flag or the observation that a series contains zeros. The engine encodes part of that discipline — it *tests* Croston on every intermittent line but *selects* it only if it actually wins the backtest — but the backtest judges the method on six weeks, and on those six weeks the verdict depends as much on where the window falls in the season as on the structure of the series. Reporting how Croston fares, and why, is more informative than either hiding its defeats or celebrating its one win.

6.4 Robust local-level models win where the selection window is flat

Eight series choose moving average or SES. Five of them — the DTC hero serum, bestseller tote and winter jacket, and the manufacturer's flagship hub and premium Pro — are not weakly seasonal lines at all: by construction they carry strong Q4 or winter seasonality (monthly multipliers spanning 0.55–1.70, against week-to-week noise of ± 11 –14%). What they share is the selection window. The six held-out weeks (12 May–16 June 2025) fall in the flat, near-average stretch of these Q4-peaking profiles (the serum's multipliers are 1.00 and 1.02 in May and June, the hub's 1.02 and 1.05) or, for the jacket, in its summer trough, where a three-week average tracks the level three steps ahead as well as a three-parameter seasonal model and with less variance. This is still the bias-variance trade-off, but made operational by the window rather than by the series: at the thirteen-week operational horizon with four origins Holt-Winters overtakes the local-level models on the serum and hub, and regression does so on the tote and jacket; only the Pro keeps its moving average. The sixth, the volume sensor, is the one genuinely weakly seasonal line (multipliers 0.90–1.08), though it too goes to Holt-Winters at thirteen weeks. The remaining two are explained by steps. The SaaS enterprise segment grows steadily but carries one jump, the large logo that lands in November 2024, which Holt's trend reads as acceleration and over-projects; SES, which ignores trend, lags by roughly a month's growth and is the smaller miss three months ahead. The academy's Leadership course is neither high-volume nor noisy: a corporate framework agreement adds about 45 students a month from January 2025, the first month of the hold-out, and no candidate models the step, which is why its WAPE (27.8%) is the highest outside the intermittent

lines and why SES leads Holt by only 0.55 points.

6.5 Regression earns its place on structured trends, including decline

Five series choose OLS regression on trend and seasonal dummies, and the pattern is coherent: the brewery’s fast-growing low/no line, the DTC candle launch, the steady UX Design Diploma and — notably — two *declining* lines, the manufacturer’s end-of-life Hub v1 (−12%/yr) and the academy’s shrinking Business Foundations course (−6%/yr). A linear trend-plus-season regression extrapolates a clean, monotone trajectory more crisply than a smoother that must infer the trend from recent levels, and handles decline as naturally as growth. The 53-column weekly design is not, on this evidence, the liability it might seem: regression is runner-up on nine further series and becomes the most frequent champion when the horizon lengthens (Table 4). Its one apparent failure — on the lager and serum the driver-augmented regression is the *worst* candidate — comes from how the engine backtests it (§3.2, §5.3). Scored with the held-out periods’ driver values, it improves from 28.7% to 14.4% on the lager, from 15.2% to 6.8% on the IPA, from 43.5% to 10.3% on the serum and from 16.6% to 7.8% on the bootcamp, which would beat Holt-Winters (9.9%) there; on the hub the substitution flatters it instead (5.1% against 6.1%). Regression is powerful where trend is the dominant, clean signal; whether its drivers add anything cannot be judged by this backtest until the engine passes it the held-out driver values, because a backtest can discriminate only between the models it actually runs.

6.6 Why the operational horizon must match the domain

The horizon and bucket are not free statistical parameters; they are dictated by the decision the forecast serves. The three inventory domains must answer “*how much do I order, and when, to avoid a stock-out before the next delivery?*” Supplier lead times are 4–8 weeks, review is weekly, and the binding planning window is therefore a quarter of weekly buckets — hence $s=52$ and a 13-week horizon. Aggregating these businesses to monthly buckets would destroy the resolution needed to time a reorder against a five-week lead time and would blur the very weekly promotional and weather spikes that drive stock-outs. The two non-inventory domains answer different questions — “*what will MRR and runway be next quarter?*” and “*will the September intake oversubscribe tutor capacity?*” — whose natural unit is the month; recurring revenue is billed monthly and enrolment is termly, so a weekly bucket would invent structure that has no operational referent and no data to support it. Matching granularity to cadence is thus simultaneously an accuracy decision (the model sees the right signal-to-noise) and an interpretability decision (the forecast lands in the operator’s decision unit). A single engine can serve all five only because granularity and horizon are configuration, not hard-coded assumptions — the same code runs at $s=52$, $H=13$ and at $s=12$, $H=12$ without modification.

A subtler design point concerns the *selection* horizon: champions are chosen on a three-step-ahead backtest but deployed to 13- or 12-step forecasts. This favours near-term accuracy but does not optimise the full operational horizon, and Table 4 shows how often the two disagree: 16 of the 31 champions change when selection is run at the operational horizon. The argument of this section therefore applies to selection as much as to forecasting, and we flag the current design as a limitation (§7).

7. Limitations

Several limitations bound the claims above.

1. **Semi-synthetic data.** The benchmark series are simulation anchored to public seasonal shapes, not real proprietary operational data. The reported WAPE values are genuine engine outputs under known generating processes, but they demonstrate *selection behaviour*, not field accuracy. Absolute error levels (many single-digit WAPes) are lower than one should expect on real, messier series and reflect the controlled noise of the generator. The champion assignments are reproducible but not robust: they rest on six periods per series, 17 of the 31 are narrow leads by the engine’s own test, and between 13 and 20 change when only the backtest configuration changes (§5.4). The transferable finding is the procedure — per-series selection reported with its margins and its sensitivity — more than any single assignment.
2. **Selection horizon shorter than operational horizon.** As noted, selection uses $H=3$ while forecasts are issued at 13 or 12 steps, and with four origins one period apart every series is scored on the same six periods (the weeks beginning 12 May to 16 June 2025 in the weekly domains, January to June 2025 in the monthly ones), the middle periods counted two or three times. The four-fold, recent-origin design therefore weights the most recent weeks or months heavily, which is appropriate for drift-responsiveness but samples a single stretch of the seasonal cycle. Table 4 shows the consequence: at the operational horizon 16 of the 31 champions change. Selecting at the operational horizon, or over more origins spread across the seasonal cycle, would make the assignments more stable; in the monthly domains it would also need enough history for Holt-Winters to be fitted at every origin.
3. **Classical Croston without SBA.** The engine implements classical Croston, not the Syntetos-Boylan bias-corrected estimator, and fixes its smoothing parameter at 0.1. On random intermittent demand — present here in the in-season weeks of the saison and the gift set and, without zeros, in the spares — SBA would be the stronger comparator. The engine also stores Croston’s in-sample fitted values after updating with each observation rather than before it, so the residuals behind a Croston champion’s interval are not one-step-ahead errors; for the saison the two give almost the same residual standard deviation (39.3 and 39.0 cases).
4. **Additive Holt-Winters only.** The seasonal smoother is additive; multiplicative seasonality (where seasonal amplitude scales with level) is not fitted, and the smoothers are not selected by likelihood/AIC as in a full ETS implementation. Some series with amplitude-scaling seasonality might be better served by multiplicative Holt-Winters or by log-transform-then-additive.
5. **How the backtest treats drivers.** The engine backtests the driver-augmented regression with each driver held at its full-history mean for every held-out step, because the selection routine passes it no future driver values; at deployment the same model receives the planned values. The backtest therefore scores a different model from the one deployed, and the error runs both ways (§6.5). The synthetic drivers compound the problem: they are almost deterministic functions of the calendar (R^2 of 0.96 to 1.00 on the model’s own trend and seasonal columns for temperature, ad-spend, PMI and the academy’s marketing index), so the fitted model splits each seasonal pattern between the dummies and the drivers, and the mean then removes the drivers’ share. Fourier-term seasonality would cure neither cause; passing the held-out driver values to the backtest, and using drivers that carry information beyond the calendar, would.
6. **No probabilistic or hierarchical evaluation.** Accuracy is assessed on point forecasts via WAPE and MAE; the prediction intervals use a fixed widening factor (§3.1), and their calibration and coverage are computed by the engine but not evaluated here; no cross-series hierarchical reconciliation (as in M5) is attempted. Metrics also degenerate on the most intermittent windows: WAPE is undefined when all held-out actuals are zero, and the earlier sMAPE fallback decided the saison by scoring every non-zero forecast

against a zero actual at its maximum (\$5.3). The MAE fallback avoids that, but a scaled measure such as MASE would allow comparison across series.

7. **Single reference implementation.** Results reflect one engine’s parameter grids and defaults. A port to a Python library such as statsmodels or StatsForecast should reproduce most champions, but different grids and defaults will shift borderline cases — and Table 2 shows that many cases here are borderline. A port should therefore be checked series by series against this engine’s outputs (Table 2) as golden values before its champions are relied on.

8. Conclusion

Across five operationally distinct domains and 31 series, a disciplined procedure — build a bank of transparent classical models, backtest each on rolling origins, and select the per-series champion by held-out error — produces an interpretable map from series structure to model choice at the three-step selection horizon. Strong annual seasonality with sufficient history selects additive Holt-Winters; compounding recurring revenue with no season selects Holt’s trend; clean monotone trends, including decline, select regression; series whose hold-out window falls in a flat stretch of the season select robust local-level smoothers; and intermittency divides three ways — scheduled zeros select seasonal-naive, random in-season arrivals select Croston on the hold-out the engine uses, and random batches on a baseline reduce Croston to exponential smoothing. The map is also fragile, and saying so is part of the method: the assignments rest on six periods per series, most are narrow leads, and between 13 and 20 of the 31 change when only the backtest’s horizon or number of origins changes. The practical thesis is twofold: first, that no single method is best, and the honest way to exploit that is to let each series declare its model through out-of-sample backtesting rather than by analyst default, reporting the margin and the sensitivity alongside the choice; and second, that the forecasting horizon and bucket are operational commitments — they must match the domain’s decision cadence and lead-time structure, so that the model sees the right signal and the forecast lands in the operator’s decision unit — and the evidence here is that the selection horizon should match them too. A single dependency-free engine can serve a brewery, an online retailer, a hardware manufacturer, a SaaS company and a training provider precisely because these choices are configuration, and because selection is evidence, not opinion.

Reproducibility statement

All figures in this version were produced by executing the reference engine as published on the author’s site since 30 September 2026 (forecast-engine.js, at <https://civfarmer.github.io/projects/vantage/engine/forecast-engine.js>; the SHA-256 checksum of that file begins 590f6f4f44034bce) on the five benchmark datasets published beside it (brewery-data.js, dtc-data.js, manufacturing-data.js, saas-data.js, education-data.js, unchanged since July 2026) under the protocol of §3.2 (rolling-origin backtest, $H=3$, four expanding-window folds, ranked by MAE with RMSE as tie-break, WAPE reported where defined) and the operational-horizon forecast of §3.1; the sensitivity runs of §5.4 and the driver re-scoring of §5.3 change only the settings stated there. Because the datasets are deterministically seeded, the evaluation is exactly reproducible. Version 1.0 of this paper (July 2026) was computed with the engine as it stood before 30 September 2026, which ranked by WAPE with an sMAPE fallback; with that rule the engine reproduces the Version 1.0 figures exactly, and the two versions’ results differ only in the Solstice Saison’s champion and in the metric shown for the two lines without a WAPE. The datasets are semi-synthetic —

seasonal shapes anchored to the public sources in Table 1, operational detail clearly-labelled simulation.

References

- Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192–213. <https://doi.org/10.1016/j.ins.2011.12.028>
- Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, 23(3), 289–303. <https://doi.org/10.1057/jors.1972.50>
- Gardner, E. S., Jr. (2006). Exponential smoothing: The state of the art — Part II. *International Journal of Forecasting*, 22(4), 637–666. <https://doi.org/10.1016/j.ijforecast.2006.03.005>
- Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. Office of Naval Research Memorandum 52, Carnegie Institute of Technology, Pittsburgh. Reprinted in *International Journal of Forecasting* (2004), 20(1), 5–10. <https://doi.org/10.1016/j.ijforecast.2003.09.015>
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts, Melbourne. Retrieved 16 October 2026, from <https://otexts.com/fpp3/>
- 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>
- 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](https://doi.org/10.1016/S0169-2070(01)00110-8)
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4), 1346–1364. <https://doi.org/10.1016/j.ijforecast.2021.11.013>
- Ord, J. K., Koehler, A. B., & Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, 92(440), 1621–1629. <https://doi.org/10.1080/01621459.1997.10473684>
- Syntetos, A. A., & Boylan, J. E. (2001). On the bias of intermittent demand estimates. *International Journal of Production Economics*, 71(1–3), 457–466. [https://doi.org/10.1016/S0925-5273\(00\)00143-2](https://doi.org/10.1016/S0925-5273(00)00143-2)
- Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting*, 21(2), 303–314. <https://doi.org/10.1016/j.ijforecast.2004.10.001>
- Syntetos, A. A., Boylan, J. E., & Croston, J. D. (2005). On the categorization of demand patterns. *Journal of the Operational Research Society*, 56(5), 495–503. <https://doi.org/10.1057/palgrave.jors.2601841>
- Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437–450. [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)
- Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3), 324–342. <https://doi.org/10.1287/mnsc.6.3.324>
-

Further reading and points of disagreement (July 2026 addendum)

Two strands of the forecast-evaluation literature bear directly on the design choices here and should be named. First, the scoring-rule literature is the canonical answer to the question this paper’s backtesting protocol raises implicitly: a point-error metric rewards a particular functional of the forecast distribution — absolute-error measures such as MASE the median, squared-error measures the mean (Gneiting, 2011) — and a system that issues intervals should ultimately be scored with proper scoring rules (interval or CRPS scoring) rather than point metrics alone (Gneiting and Raftery, 2007) — an upgrade path this framework accommodates. Second, the M-competition record, notably the M4 findings, is a standing caution against method complexity: simple and combined statistical baselines remain hard to beat, which is precisely why this paper benchmarks against seasonal-naïve and moving-average baselines before anything else. The disagreement to acknowledge is between in-sample sophistication and out-of-sample humility; the design lands with the latter, and the cited canon is the reason.

- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762. <https://doi.org/10.1198/jasa.2011.r10138>
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74. <https://doi.org/10.1016/j.ijforecast.2019.04.014>