Perishability to Predictability: Advanced Demand Planning for Food Supply Chains

by

Utkarsh Goel

and

Sharvil Dinesh Khot

Submitted on May 8th, 2026, in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management
at the Massachusetts Institute of Technology

ABSTRACT

Isla Délice, a leading French halal food manufacturer, faces significant demand-planning challenges across a portfolio of 424 active SKUs, with diverse demand patterns ranging from high-volume, stable products to low-volume, intermittent items. The current forecasting approach lacks the granularity and statistical rigor needed to minimize forecast error, leading to suboptimal inventory levels, service failures, and excess safety-stock costs. This capstone project develops a segmentation-driven demand forecasting framework. SKUs are first clustered by demand behavior using K-means on Average Demand Interval (ADI) and squared Coefficient of Variation (CV^2). For each cluster, the best-performing forecasting model is selected from a candidate set that includes exponential smoothing methods, AutoSARIMA, the Syntetos-Boylan Approximation (SBA), Prophet, and LightGBM. Results from cluster-level time-series modeling indicate strong separability into four demand pattern types (Smooth, Intermittent, Erratic, Lumpy), each with a distinct best-performing forecasting model. The framework produces transparent, cluster-based recommendations that planners can interpret and apply alongside their existing ABC-XYZ view, with cluster-winning models leading to meaningful WMAPE reductions in the Intermittent and Smooth clusters (which together cover roughly two-thirds of active SKUs) relative to a uniform forecasting strategy.

Capstone Advisor: Dr. Inma Borrella

Title: Research Scientist

ACKNOWLEDGEMENTS

The authors would like to thank Benoit Lescarret, President of Opagan, and, Olivier Brandstatt, Senior Advisor at Opagan, for their guidance, domain expertise, and continued engagement throughout this project. Their practitioner perspective on supply chain planning was invaluable in shaping both the framing and execution of this work. We would also like to thank Dr. Inma Borrella, our capstone advisor, for her academic guidance, thoughtful feedback, and support throughout the project.

TABLE OF CONTENTS

1. INTRODUCTION

The global food manufacturing sector operates amid complex, volatile demand, shaped by shifting consumer preferences, seasonal cycles, promotional activity, and external events. Volatility propagates through the supply chain: poor forecasts drive stockouts, overstocking, expedited production runs, and degraded service levels, making accurate demand forecasts a critical competitive differentiator. The challenge is acute for companies managing hundreds of Stock Keeping Units (SKUs) across multiple markets and demand regimes.

The sponsor of this project, Opagan, is a French Supply chain Consulting firm that is developing a proprietary ERP tool called SC Control to help companies of various sizes across multiple industries tackle their supply chain challenges.

Isla Délice, one of France's leading halal food companies and a flagship client of the project's sponsor, Opagan, serves as a representative case study of these demand-forecasting challenges. With an active portfolio of 424 SKUs spanning categories from processed meats to dairy alternatives, Isla Délice operates in a fast-moving consumer goods (FMCG) environment where stockouts and overstock events carry direct financial and reputational consequences. The planning team has recognized that its current process relies heavily on planners' judgment and aggregate-level forecasts that do not account for heterogeneous demand characteristics across the SKU portfolio. Certain high-volume, event-driven SKUs, particularly those tied to seasonal peaks such as Ramadan, are frequently under-forecasted. In contrast, a significant share of long-tail, low-velocity SKUs consume disproportionate planning resources despite generating minimal revenue.

1.1 Problem Statement and Goals

The central problem this capstone addresses is how to redesign Isla Délice's demand forecasting process to account for heterogeneous demand patterns across its SKU portfolio, thereby reducing forecast error and improving planning efficiency. The current one-size-fits-all approach systematically misallocates planning resources and underperforms on both high-value variable SKUs and stable niche products.

Three sub-questions follow:

1. How can active SKUs be meaningfully segmented based on historical demand patterns?
2. Which forecasting models are best suited to each demand segment, and how should model assignment be operationalized?
3. How can the resulting framework be integrated into Opagan's Software as a Service (SaaS) planning environment and sustained by Isla Délice's planning team with minimal external dependency?

The project delivers four outputs: (1) a validated SKU segmentation taxonomy with cluster profiles and business interpretation; (2) a model assignment matrix linking each demand segment to its recommended forecasting approach; (3) forecast outputs for a representative 113-SKU sample with accuracy benchmarking; and (4) an integration roadmap for adoption within Opagan's planning platform. Beyond the immediate sponsor deliverables, the work demonstrates how K-means clustering on ADI and CV^2 can group SKUs by demand behavior, and how assigning tailored forecasting models to each group improves performance in a real-world

FMCG setting, with a framework that is transferable to other food manufacturers facing similar portfolio complexity.

2. STATE OF THE PRACTICE

This section synthesizes the streams of literature that informed the methodology, organized into demand classification and SKU segmentation, time-series forecasting methods, and the application of machine learning to FMCG demand forecasting.

2.1 Demand Classification and SKU Segmentation

A foundational challenge in demand planning is the heterogeneity of demand patterns across SKU portfolios. Croston (1972) introduced the statistical basis for forecasting intermittent items by separately modeling demand size and demand interval. Syntetos et al. (2005) later proposed a two-dimensional classification based on the Average Demand Interval (ADI) and the squared Coefficient of Variation (CV^2), distinguishing Smooth, Intermittent, Erratic, and Lumpy demand patterns. These approaches have been validated across FMCG, spare parts, and pharmaceuticals (Boylan & Syntetos, 2007).

Traditional ABC-XYZ segmentation, which classifies SKUs by demand volume and variability, remains the dominant industry framework but relies on manually defined thresholds and does not adapt to distributional nuances within clusters. Several studies highlight that ABC-XYZ segments can be internally heterogeneous, particularly in the Z-category, where demand intermittency and lumpiness are conflated (Huiskonen, 2001). This limitation motivated the project's pivot toward a data-driven K-means clustering approach (Section 3.3.1).

K-means partitions observations into a pre-specified number of clusters by minimizing within-cluster variance around centroids (MacQueen, 1967). It is widely used for SKU classification because of its computational scalability and the interpretability of its centroid-based output. Hierarchical methods scale poorly to large portfolios; density-based methods (DBSCAN) require tuning of density parameters that are difficult to justify for ADI/CV^2 feature spaces. K-means was therefore selected, with the choice of k determined empirically via the silhouette coefficient.

2.2 Time-Series Forecasting Models

Statistical forecasting literature offers a rich toolkit. Exponential smoothing methods (Simple Exponential Smoothing, Holt's linear trend, and Holt-Winters' seasonal extension) are widely used for their computational efficiency and adaptability (Hyndman & Athanasopoulos, 2021). AutoRegressive Integrated Moving Average (ARIMA) and its seasonal variant SARIMA capture more complex autocorrelation structures and are effective for high-volume SKUs with consistent seasonal patterns (Box & Jenkins, 1976). Prophet (Taylor & Letham, 2018) decomposes a series into a piecewise-linear trend, a Fourier-series seasonal component, and user-specified holiday or event effects, with parameters estimated via Bayesian inference; it is well-suited to settings with exogenous regressors such as Ramadan windows or promotional events.

For intermittent demand, Croston's method and its bias-corrected variant, the Syntetos-Boylan Approximation (SBA), remain the established benchmarks, modeling demand intervals and sizes separately to produce calibrated forecasts for sparse series.

2.3 Machine Learning in FMCG Forecasting

ML applications in demand forecasting have accelerated over the past decade. Gradient boosting (XGBoost, LightGBM) and neural architectures have demonstrated superior accuracy on large, feature-rich datasets (Makridakis et al., 2020). However, Fildes et al. (2022) caution that ML models require substantial historical data, careful feature engineering, and ongoing maintenance. Specifically, in FMCG, hybrid approaches that combine a statistical baseline with ML-based adjustments for promotional and causal effects show the most consistent improvements (Huang et al., 2020).

This literature shaped two architectural choices. First, LightGBM is included among five candidate models in the cluster-level tournament, competing head-to-head on WMAPE rather than being layered on top of a statistical baseline. Second, known exogenous drivers, specifically the pre-Ramadan demand uplift observed in Isla Délice's data, are incorporated directly as external regressors within AutoSARIMA and Prophet. Promotional effects are handled upstream in data preparation: promotion-driven weeks are flagged, and the promotional signal is cleaned before forecasting (Section 3.2). Together, these literature streams motivate the project's central design choice: clustering SKUs by demand behavior and selecting the best-performing model for each cluster, rather than applying a single model uniformly across the portfolio.

3. METHODOLOGY

The methodology proceeds through three stages: methodology selection, data preparation, and application.

3.1 Methodology Selection

The core decision was whether to apply a uniform forecasting model across all SKUs or adopt a segmentation-first approach, assigning models based on demand patterns. The literature establishes that no single model dominates across all demand types; the optimal model varies systematically with variability and intermittency. This evidence supports a segmentation-first architecture.

Within segmentation approaches, ABC-XYZ was considered first due to its prevalence in Opagan's planning vocabulary. However, ABC-XYZ has known limitations: threshold sensitivity, intra-segment heterogeneity, and an inability to distinguish Erratic from Lumpy demand within the Z-category. The project therefore adopted K-means clustering on ADI and CV^2 features, which is data-driven, objective, and scalable to the full active portfolio. The optimal k was selected empirically using the silhouette coefficient (S-score), which measures how well each observation fits within its assigned cluster relative to neighbors. K-means was fitted for k from 3 to 10; the value that maximized the S-score (k=4, S-score = 0.478; see Figure 1) was selected, corresponding to the four canonical demand pattern types in the Syntetos-Boylan framework. A detailed comparison of these methods is presented in Table 1.

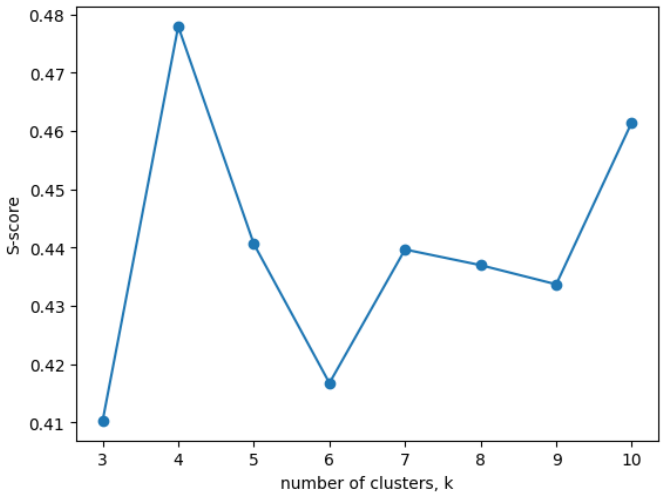


Figure 1. Silhouette score (S-score) v.s number of clusters k.

Criterion	**ABC-XYZ**	**K-means on ADI/CV²**	**Inference**
Theoretical basis	Revenue (A/B/C) × variability (X/Y/Z); manual thresholds	ADI & CV² + minimization of within-cluster variance	K-means provides an objective, statistically grounded classification
Cluster definition	Fixed 3×3 matrix; analyst judgment sets boundaries	Data-driven k=4 from S-score optimization	No arbitrary threshold setting required
Pattern types captured	Smooth, variable, Z-class (conflates erratic/intermittent)	Smooth, Intermittent, Erratic, Lumpy - distinct	Separating Erratic from Lumpy enables differentiated model assignment
Intra-segment homogeneity	Low-to-moderate, especially C-Z	High; minimizes within-cluster variance	More reliable model assignment within each group
Adaptability	Static unless re-segmented manually	Dynamic; clusters retrainable as patterns evolve	Quarterly re-clustering catches SKU migration

Table 1. Comparison of segmentation approaches

3.2 Data Collection and Preparation

The dataset, provided by Isla Délice, comprises 48 months of historical weekly sales transactions at the SKU × customer cluster and SKU × country level for the full active portfolio of 424 SKUs across five product categories. An additional 149 D-class SKUs (no recorded sales in the prior two years) were excluded to avoid distorting segmentation statistics.

Data preparation followed four steps:

1. Raw transaction records were cleaned to remove duplicates, correct unit-of-measure inconsistencies, and address missing values via forward-fill imputation for gaps of up to two consecutive periods.

2. Demand series were reconstructed at weekly granularity, with zero-demand weeks explicitly encoded rather than treated as missing, which is a critical step for accurate ADI computation.

3. Average Demand Interval (ADI) and the Squared Coefficient of Variation (CV^2) were calculated for each SKU across the full historical window: $ADI = N / N_{nz}$, where N is the total number of observed weeks and N_{nz} is the number of weeks with positive demand; and $CV^2 = (\sigma / \mu)^2$ is, where σ is the standard deviation of weekly demand and μ is the mean weekly demand, both computed on the raw weekly demand series for each SKU. Both statistics were then transformed using a Yeo-Johnson power transformation (a generalization of the Box-Cox transform that reshapes skewed distributions toward normality and accommodates zero and negative values) and standardized to ensure long tails did not dominate K-means distances.

4. Promotional weeks were flagged using Isla Délice's promotional calendar, and weekly demand for a customer during promotional weeks was replaced with the rolling median of surrounding non-promotion weeks (a 4-week window on each side; promotion-week neighbors excluded from the median). This produces a baseline demand series that the forecasting models see as if no promotion had occurred; promotional uplift is layered back in downstream during the planner workflow (Section 5.2).

In addition, a pre-Ramadan exogenous regressor was constructed, with a binary flag set to 1 for the two-week window preceding each Ramadan period in the historical demand data. This is also extended to forecast horizons and used as an external regressor within AutoSARIMA and Prophet to directly capture the pre-Ramadan demand uplift.

The four panels in Figure 2 illustrate how promotional-week imputation works in practice. For each SKU, orange shading marks weeks flagged as promotional; the red line shows raw demand during those weeks, inflated by the promotional uplift; and the black line shows the cleaned baseline after rolling-median imputation replaces those values. Where red and black diverge, the model sees a smoother, promotion-free signal, most visibly in SKU 1PT100000.01, where nearly continuous promotional activity would otherwise dominate the training data, and in SKU 1DD306300, where the imputation corrects a short but steep promo-driven ramp early in the series.

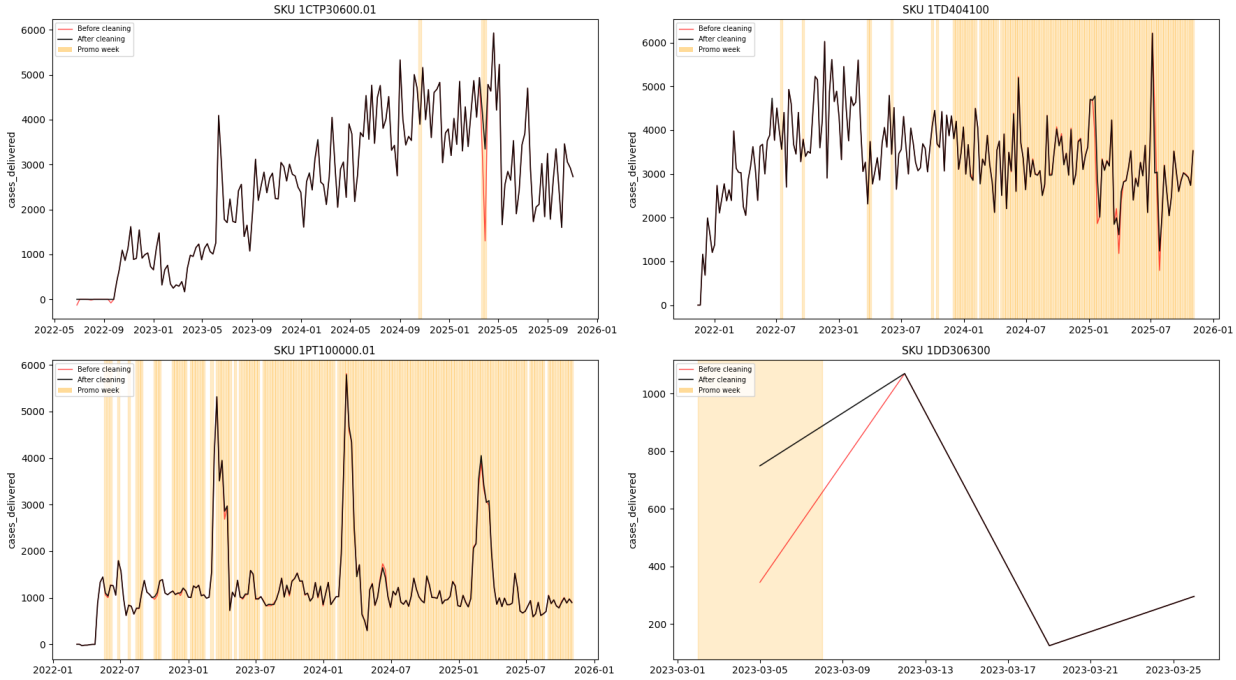


Figure 2: Effect of promotional-week imputation for the four SKUs with the largest demand corrections

3.3 Application of Methodology

The application proceeded through three phases: SKU clustering, stratified sampling, and model assignment.

Phase 1: SKU Clustering via K-means. K-means with k=4 was applied to the standardized ADI and CV^2 feature space for the 424 active SKUs. The four resulting clusters were labeled by demand pattern based on their position in the feature space:

- Cluster 0 (Intermittent): high ADI, low CV^2, 125 SKUs
- Cluster 1 (Smooth): low ADI, low CV^2, 188 SKUs
- Cluster 2 (Erratic): low–moderate ADI, high CV^2, 88 SKUs
- Cluster 3 (Lumpy): high ADI, high CV^2, 23 SKUs. (Note: only ~5 of these 23 points appear visible in Figure 3 because the x-axis is truncated for readability; the remaining Cluster 3 SKUs sit at CV values beyond the chart's right edge.)

The cluster distribution is shown in Figure 3

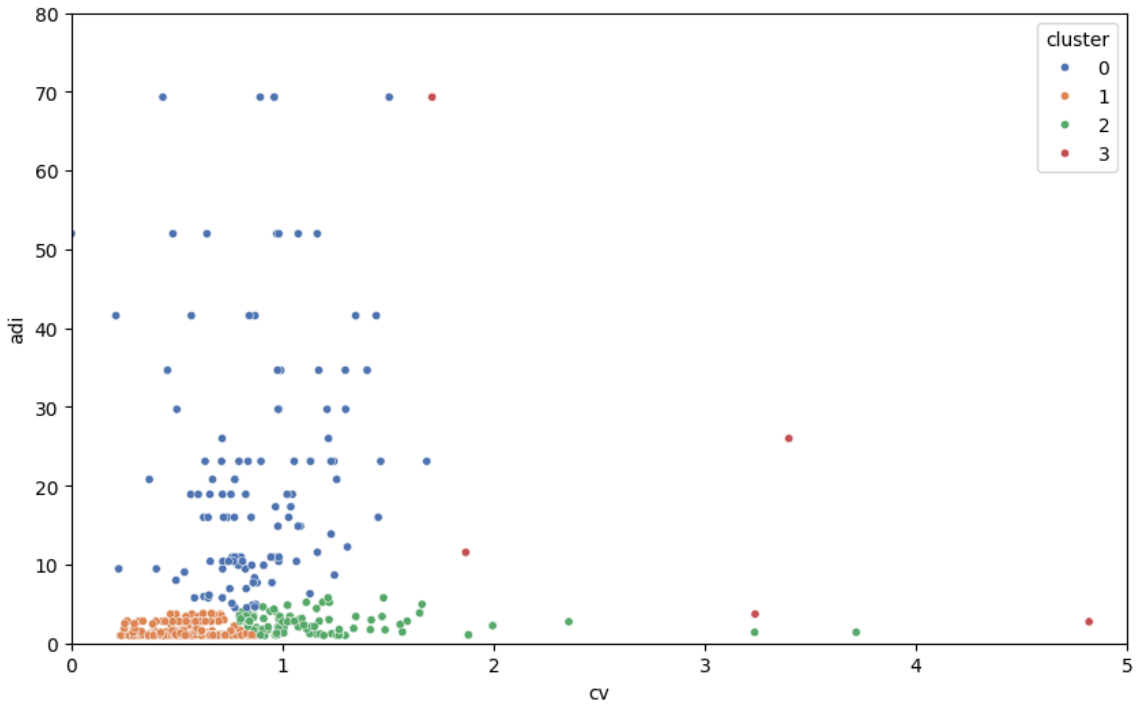


Figure 3: K-means cluster assignments in the ADI–CV feature space.

Each color in Figure 3 corresponds to one of the four clusters: blue for Cluster 0 (Intermittent, high ADI), orange for Cluster 1 (Smooth, low ADI and low CV), green for Cluster 2 (Erratic, elevated CV), and red for Cluster 3 (Lumpy, high ADI and high CV). The x-axis displays CV; clustering itself was performed on CV^2 per Section 3.2.

The pattern labels assigned to each cluster (Smooth, Intermittent, Erratic, Lumpy) follow the canonical Syntetos-Boylan vocabulary and are intended as interpretive aids for cross-referencing this analysis with the broader demand-classification literature and with planners' existing mental models. They are not validated descriptors of the underlying behavior. In particular, the Smooth cluster contains SKUs that exhibit substantial week-to-week variability despite their relatively low ADI and CV^2 values; "Smooth" denotes only that these SKUs sit in the low-ADI, low-CV^2 region of the feature space relative to the rest of the active portfolio, not that their demand is genuinely flat or low-noise in absolute terms. The same caveat applies to the other three labels.

Phase 2: Stratified Cluster Sampling. To make model development tractable while preserving representativeness, 30 SKUs per cluster were sampled (the entire 23-SKU Lumpy cluster was retained, total sample size: 113). Within each cluster, six near-centroid SKUs (most representative of the cluster average) and six boundary SKUs (outliers) were selected via distance-based ranking. The remaining 15 SKUs were drawn via proportional random sampling across ADI/CV^2 quartiles to capture within-cluster variation. A forced-inclusion constraint required that at least the top three SKUs by total sales volume within each cluster be included, ensuring that business-critical items were represented regardless of their statistical positioning.

Phase 3: Forecasting Model Assignment. For each sampled SKU, weekly demand was evaluated using a walk-forward expanding-window cross-validation design with three origins per SKU. The test window was capped at 13 weeks to reflect a realistic operational planning

horizon; origins were spaced evenly between 50% of the series length and the most recent feasible cutoff, with all training sets expanding as origins advance. The candidate set was: exponential smoothing methods (Holt-Winters with additive trend and additive seasonality when the SKU had ≥104 weeks of history; otherwise Holt's linear-trend method, with Simple Exponential Smoothing as the final fallback); AutoSARIMA (with the pre-Ramadan exogenous regressor); SBA (the bias-corrected Croston variant designed for intermittent demand); Prophet (with the pre-Ramadan regressor); and LightGBM, with optional Seasonal-Trend decomposition using Loess (STL) pre-whitening when sufficient seasonal history was available (under STL pre-whitening, the trend and seasonal components are removed before fitting, the model learns the residual, and the components are added back to produce the final forecast).

Forecast accuracy was evaluated using Weighted Mean Absolute Percentage Error (WMAPE), defined as $\text{WMAPE} = \Sigma|y_t - \hat{y}_t| / \Sigma|y_t|$, where y_t is the actual demand and $\hat{y}_t$ is the forecast for week t over the test horizon. This formulation gives higher-demand weeks greater influence on the SKU-level error. Per-SKU WMAPE was averaged across the three cross-validation origins, then aggregated to the cluster level by weighing each SKU's WMAPE by its total demand volume, so high-volume SKUs drive the cluster-level score more than long-tail items. Within each cluster, the top 10% of SKUs by best-model WMAPE were excluded as outliers, and SKUs with no test-period demand were dropped because WMAPE is undefined for them. The resulting cluster-level rankings, which indicate which model performs best over a 13-week test window, were then used to generate production forecasts for the full active portfolio over a longer six-month deployment horizon (26 weekly periods).

In addition to point forecasts, each model produces a 95% prediction interval for the test and future horizon periods. Interval construction is model-native where possible. AutoSARIMA uses StatsForecast's analytic confidence bounds; Prophet uses its built-in Bayesian uncertainty estimate with interval_width=0.95; ExpSmooth uses simulation with 200 repetitions on the fitted state-space model; LightGBM produces separate quantile regressors at the 2.5th and 97.5th percentiles; SBA, which has no analytic confidence interval, derives bounds from the standard deviation of in-sample residuals assuming a Gaussian error distribution. The prediction intervals are stored alongside point forecasts in the output and are intended for downstream use in safety stock sizing and exception flagging, not for model selection.

4. RESULTS

This section presents the empirical results of the segmentation and forecasting framework. Section 4.1 reports the four-cluster structure produced by K-means on the ADI/CV² features and the per-cluster performance of the five candidate forecasting models, identifying the best-performing model for each cluster and documenting how those choices behave when applied to the full active portfolio. Section 4.2 outlines the limitations that bound the interpretation of these results.

4.1 Forecasting Model Performance

After removing the top 10% of SKUs by best-model WMAPE within each cluster as outliers and dropping SKUs with zero test-period demand, the final evaluation set comprised 81 SKUs across the four active clusters: 23 Intermittent, 27 Smooth, 25 Erratic, and 6 Lumpy. The Lumpy cluster’s small evaluation set reflects the high attrition from the outlier and zero-demand filters

on a cluster that began with only 23 SKUs; results for this cluster should be read as indicative rather than definitive.

Table 2 reports the demand-weighted WMAPE for each candidate model within each cluster, along with the best-performing model. Cluster-level WMAPE was evaluated over a 13-week (3-month) test window per cross-validation origin, reflecting Isla Délice's S&OP planning horizon. The longer 26-week (six-month) horizon is reserved for production deployment forecasts (Section 4.1).

Table 2: Demand-weighted WMAPE per model per cluster (lower is better; bold = winner)

Cluster	Name	SKUs (eval)	ExpSmooth	Auto-SARIMA	SBA	Prophet	LightGBM	Best
0	Intermittent	23	148%	1,292%	2,560%	4,074%	**107%**	LightGBM
1	Smooth	27	121%	**21%**	203%	36%	233%	AutoSARIMA
2	Erratic	25	**113%**	203%	257%	182%	179%	ExpSmooth
3	Lumpy	6	**302%**	4,904%	1,734%	684%	1,312%	ExpSmooth

WMAPE values above 100% reflect series where total absolute forecast error exceeds total actual demand over the test window, especially common in sparse, intermittent demand patterns. The cluster-level findings are as follows:

LightGBM won cluster 0 (Intermittent) at 107% WMAPE, well ahead of ExpSmooth at 148%. This is a surprising result relative to the Syntetos-Boylan framework, which would predict SBA or Croston as the strongest performer for high-ADI series. Two factors likely contribute: the absolute volumes in this cluster are not as low as those in classic intermittent benchmarks (i.e., zero-demand weeks coexist with substantial peak demand), and STL pre-whitening allows LightGBM to model the strong Ramadan seasonality directly while gradient boosting captures the residual non-zero pattern. SBA performs poorly here (2560%) because its constant-rate forecast cannot track the seasonal peaks.

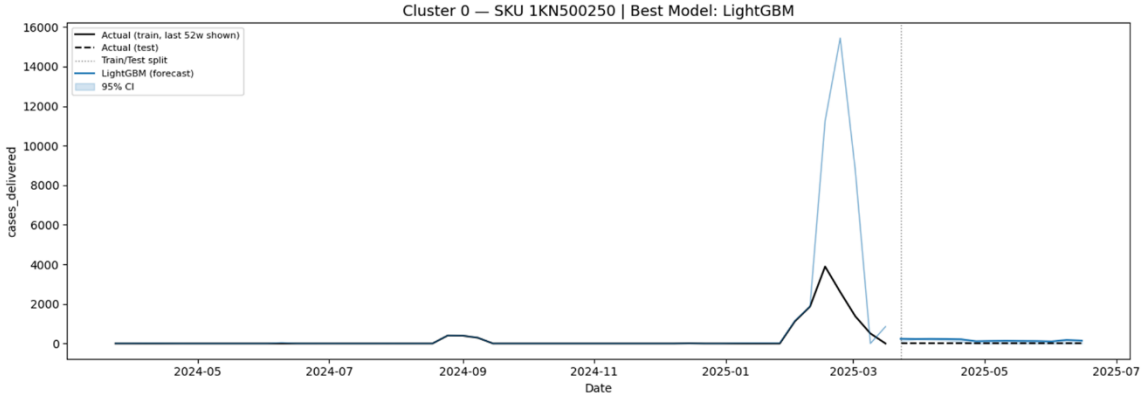
Cluster 1 (Smooth), the largest cluster with 188 active SKUs, was won by AutoSARIMA at 21% WMAPE, with ExpSmooth a distant second at 121%, and the remaining models clustered between 203% and 233%. AutoSARIMA's wide margin reflects two strengths: the pre-Ramadan exogenous regressor lets it directly model the recurring seasonal uplift in this cluster, and the autoregressive structure captures short-horizon mean reversion that flatter methods miss. The cluster's "Smooth" label here refers only to its position in the relative ADI/CV^2 feature space; in absolute terms, these series exhibit substantial week-to-week variability, which is what penalizes the constant-rate methods (SBA, ExpSmooth fallbacks).

Cluster 2 (Erratic) was won by ExpSmooth at 113% WMAPE, with LightGBM second at 179% and Prophet third at 182%. The wider best-to-worst spread (113% to 257%, ~127%), much larger than the Smooth cluster's intra-model spread once AutoSARIMA is included, indicates that model choice matters significantly here despite the absence of a dominant temporal pattern. ExpSmooth's adaptive level-tracking handles the high week-to-week variability without overfitting spurious seasonality, which the more structured methods (AutoSARIMA, Prophet) attempt to impose.

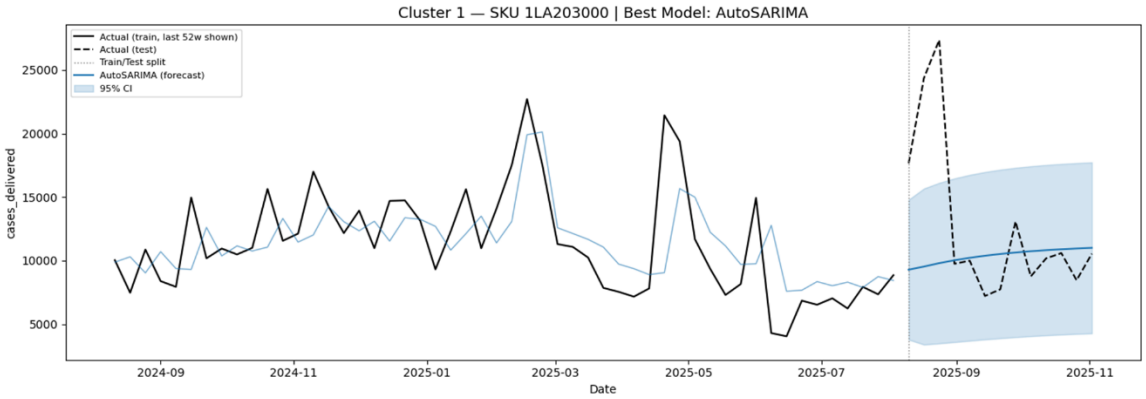
ExpSmooth won cluster 3 (Lumpy) at 302% WMAPE; with only 6 evaluation SKUs surviving the filters, this cluster's results should be treated as directional rather than conclusive. All methods produced elevated error here, particularly AutoSARIMA at 4904%, which is driven by extreme behavior on one or two outlier SKUs in this small evaluation set rather than systematic model failure. The pattern is consistent with the broader literature on the difficulty of lumpy-demand forecasting.

Figure 4 illustrates the cluster-level behavior by overlaying each candidate model's forecast against actual demand for a representative SKU in each cluster, with the train/test split marked by a dotted line. The shaded band in each panel represents the 95% prediction interval; its width varies by model and cluster, reflecting both model uncertainty and the structural difficulty of each demand pattern. The visual contrast across clusters makes the differential model performance interpretable. Intermittent-cluster SKUs combine zero-demand weeks with sharp Ramadan peaks that LightGBM (with STL pre-whitening) tracks more cleanly than constant-rate methods like SBA. Smooth-cluster SKUs exhibit relatively stable weekly demand, with multiple methods converging on similar performance. Erratic-cluster SKUs show high week-to-week variability with no dominant pattern, where all models struggle similarly. Lumpy-cluster SKUs combine intermittency with size volatility, where every method underperforms in absolute terms.

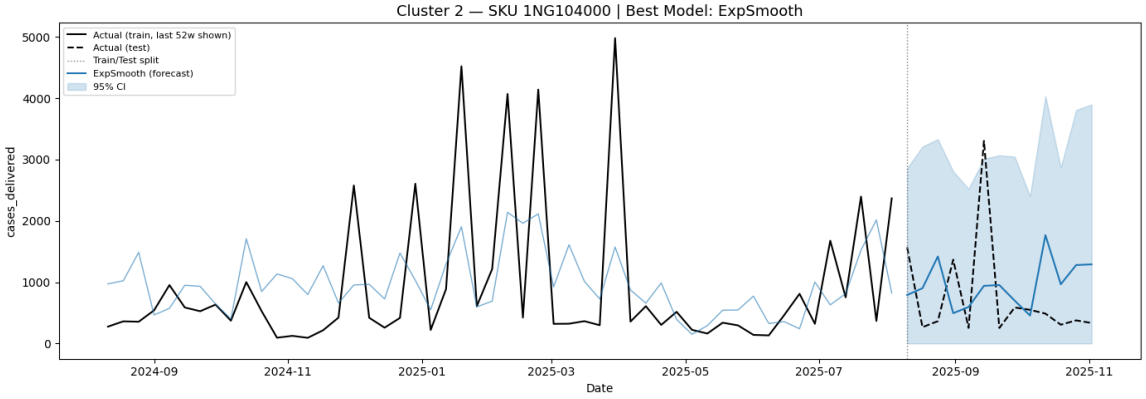
The width of the 95% prediction interval has direct operational value beyond the point forecast itself. The upper bound informs safety stock sizing for service-level targets, since it represents the demand level that will not be exceeded in 97.5% of cases. The lower-bound flags downside risk for slow-moving SKUs, where overstocking incurs shelf-life and write-off costs. The fact that intervals are wide for Erratic and Lumpy SKUs is itself the signal — these are precisely the SKUs where planners should rely on judgment-based overrides rather than automated execution (Section 5.1.2), and the bandwidth quantifies how much headroom that judgment requires.



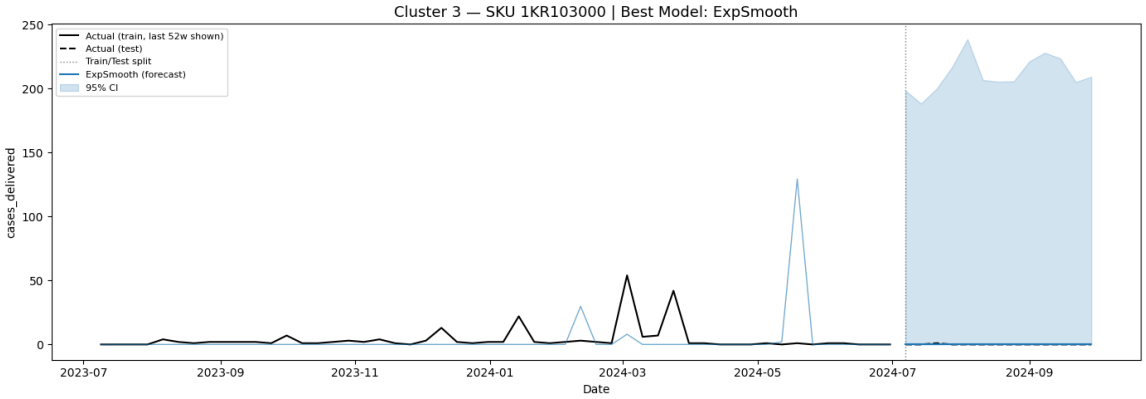
Cluster 0 (Intermittent) - LightGBM



Cluster 1 (Smooth) - AutoSARIMA



Cluster 2 (Erratic) - ExpSmooth



Cluster 3 (Lumpy) - ExpSmooth

Figure 4: Best-model forecast vs. actual demand for a representative SKU in each cluster. Solid black = training actuals; dashed black = held-out test actuals; blue line = best-model forecast; shaded blue band = 95% prediction interval; vertical dotted line = train/test split.

Three insights emerge across clusters. First, no single model is universally optimal; performance depends strongly on demand characteristics, reinforcing the project's central hypothesis that segmentation-driven model assignment is a more accurate strategy than uniform model

deployment. Second, the cluster-level approach reduces the need for SKU-by-SKU model tuning while preserving most of the accuracy benefit, which materially simplifies operational governance. Third, the winning model differs meaningfully across clusters, with LightGBM winning for Intermittent, AutoSARIMA for Smooth, ExpSmooth for both Erratic and Lumpy, and the magnitude of the gap between best and worst varies dramatically, from a 5× spread in Cluster 0 (Intermittent) to over 230× in Cluster 1 (Smooth). This heterogeneity reinforces that the choice of model is not a stylistic preference but a substantive driver of forecast accuracy in some clusters, while less critical in others.

When the cluster-winning models were applied to the full active portfolio (568 total SKUs, with 144 D-class items excluded and 28 skipped for insufficient history), 396 SKUs received a forecast. The deployment model mix was AutoSARIMA for 188 SKUs, ExpSmooth for 164, and LightGBM for 44, markedly more diverse than what a uniform-model approach would have produced and aligned with the tournament-winning models for Clusters 0 (LightGBM) and 1 (AutoSARIMA). Where the cluster's primary model fails to converge or encounters insufficient-history conditions on individual SKUs, the framework falls back to the next-ranked model; ExpSmooth's strong second- and third-place finishes across multiple clusters explain its sizeable share of the deployment, but the primary models still dominate within their respective clusters. This fallback behavior is intentional and ensures every active SKU receives a forecast. The mix raises a fair question: if multiple models share the deployment, why segment at all? Three answers. First, the within-cluster model differentiation revealed by the tournament is real and material where it matters most: in the Intermittent cluster, LightGBM beats ExpSmooth by 28%, and in the Smooth cluster, AutoSARIMA beats ExpSmooth by 83%. Second, the cluster structure provides a diagnostic layer that is independent of which model ultimately runs: assignment to Erratic or Lumpy is itself a signal to planners that the SKU needs human judgment rather than automation, regardless of the forecast value produced. Third, the cluster boundaries enable trigger-based reassignment (Section 5.1.2) so that SKUs migrating between behavior types are reclassified rather than forced into a static one-size model. The cluster framework is therefore not an alternative to a single model in deployment, but a structural layer that makes the choice of model deliberate, monitored, and reversible.

4.2 Limitations

Four structural limitations of the current analysis should be acknowledged.

First, the evaluation sample is a stratified subset of each cluster rather than the full portfolio. The sampling design is representative across centroid, boundary, and top-volume regions, but cluster-level WMAPE estimates carry sampling variance that a full-portfolio re-fit would reduce. The Lumpy cluster is most affected, with only six SKUs surviving the outlier filter; its results are directional.

Second, the clustering method embeds methodological assumptions. K-means assumes roughly spherical clusters in the feature space and requires a pre-specified k; the silhouette coefficient was used to select k=4, but a different clustering method (e.g., DBSCAN or a Gaussian mixture model) could yield different groupings.

Third, the forecasting framework is univariate in its treatment of each SKU. Prophet and AutoSARIMA support exogenous regressors, and pre-Ramadan window indicators were

included, but structured promotional calendars and pricing histories were not available in a clean form, limiting the models' ability to explain demand uplift around known drivers.

Fourth, the approach optimizes WMAPE without directly incorporating downstream inventory costs, service-level outcomes, or working-capital implications. Two SKUs with identical WMAPE can have very different cost-of-error profiles depending on lead time, shelf life, and stockout penalties; integrating a cost-aware loss function that links forecast error to inventory obsolescence (Teunter et al., 2011) would be a meaningful but separate modeling effort, scoped as future work in Section 5.2.

Fifth, model evaluation depends on the selection of the cross-validation origin. The three rolling origins are spaced between 50% of the series length and the most recent feasible cutoff, which means models with stronger long-history requirements (Holt-Winters, AutoSARIMA with seasonality) are evaluated under conditions for which they are designed. Models with shorter memory (SBA, Prophet) are not handicapped by the choice. This is a defensible design because it tests each model under the conditions it would face in deployment. However, it is not strictly model-agnostic, and a different origin scheme could shift the cluster-level rankings in clusters where the spread between top models is narrow.

5. RECOMMENDATIONS

This section presents seven recommendations for adopting cluster-based forecasting at Isla Délice (Section 5.1) and outlines priority extensions for future work (Section 5.2). The recommendations translate the analytical findings of Section 4 into operational guidance for demand planning and supply chain leadership.

5.1 Recommendations for Adoption

Each recommendation below pairs a strategic shift with the operational change needed to realize it; the recommendations are listed in approximate order of implementation priority.

- **Adopt cluster-based forecasting as the standard planning framework**

  The current forecasting process applies similar planning logic across all SKUs despite major differences in demand behavior. The results in Section 4 show that model performance varies significantly by demand type. Isla Délice should therefore adopt a cluster-based forecasting framework in which each demand cluster is assigned its best-performing forecasting model.
  The recommended default assignments for this data are:
  - LightGBM for Intermittent SKUs
  - AutoSARIMA for Smooth SKUs
  - ExpSmooth for Erratic SKUs
  - ExpSmooth for Lumpy SKUs

  The framework should also maintain a documented fallback hierarchy so that if the primary model fails to converge or lacks sufficient history, the next-best model can automatically generate a forecast.

- **Reallocate planners' effort by cluster**

The Smooth and Intermittent clusters together represent 313 active SKUs, or approximately 71% of the active portfolio. These clusters showed the strongest forecasting performance and are therefore the best candidates for automation-first planning.
Planner intervention for these SKUs should be limited to major business events such as product launches, supply disruptions, or exceptional promotional activity. The Erratic and Lumpy clusters require greater planner involvement because their demand patterns are less predictable, and statistical models alone perform less reliably. Planner judgment, commercial intelligence, and customer-level insights are therefore most valuable in these segments.

- **Refresh clusters quarterly and forecasting models monthly**

  Cluster assignment reflects the structural behavior of an SKU and should therefore be refreshed quarterly. Forecasting model parameters, however, respond to shorter-term demand dynamics and should be recalibrated monthly.
  Between scheduled refreshes, the system should flag SKUs for planner review when:
  - ADI or CV^2 changes materially
  - Forecast error exceeds the cluster average by more than 1.5 standard deviations for two consecutive cycles

  These SKUs can then be surfaced through an exception-management workflow for reassignment or manual review.

- **Onboard new SKUs through analog-based forecasting**

  New SKUs typically lack sufficient historical demand to support reliable statistical forecasting. A new SKU should therefore enter the forecasting framework only after accumulating at least 12 weeks of history and eight non-zero demand observations. Until these thresholds are reached, the system should identify the closest analog SKU within the same product family using ADI/CV^2 similarity. The new SKU can temporarily inherit the analog's cluster assignment and forecasting model while planners overlay a launch-demand profile based on commercial expectations.
  SKUs with no recorded demand over a continuous 24-month period should be reclassified as D-class items and removed from the active forecasting pool.

- **Roll out in phases and align planner KPIs to cluster-weighted WMAPE**

  Implementation should begin with a limited pilot focused on the Smooth cluster, which represents the lowest operational risk and the most stable forecasting environment. The pilot can run in parallel with the current planning process for approximately 6–8 weeks. Once the pilot stabilizes, the framework can expand sequentially to the Intermittent, Erratic, and finally Lumpy clusters.
  Planner performance metrics should also evolve alongside the forecasting process. Instead of relying on a single portfolio-wide MAPE metric, Isla Délice should adopt cluster-level WMAPE reporting. This allows performance targets to reflect the inherent difficulty of each demand type rather than treating all SKUs equally.

- **Invest in structured data capture and a quarterly demand-review forum**

  One of the largest limitations identified during the project was the lack of clean, structured data on promotions, pricing changes, and customer-specific events.
  Future improvements in forecasting will depend heavily on consistently capturing these variables at the SKU-week level. This includes:
  - Promotional calendars
  - Price changes
  - Customer-specific commercial events
  - Trade-marketing activities
  - Seasonal campaigns

  To support this process, Isla Délice should establish a quarterly demand review forum involving the sales, trade marketing, S&OP, and supply chain teams. The purpose of this review is to evaluate cluster-level forecast performance, discuss recurring forecast biases, and align on upcoming commercial events before forecasts are finalized.

5.2 Handling Promotional Effects

Promotional activity is one of the largest sources of demand distortion in the dataset. Because the forecasting models operate on a cleaned baseline demand series, promotional effects should be modeled separately and layered back into the final planner-adjusted forecast.
Three promotional effects are particularly important-

- First, promotions can create cannibalization and halo effects across related SKUs. A promotion on one product may reduce demand for substitute products while increasing demand for complementary items. Isla Délice should maintain a planner-managed SKU relationship table that identifies linked products and estimates the expected demand response during promotional periods.
- Second, promotions distort demand before and after the event itself. Customers often delay purchases before a discount period and reduce purchases immediately afterward. The current cleaning process adjusts only the promotional week, so surrounding weeks may still bias the baseline forecast. Extending the imputation window to include nearby weeks would mitigate this issue.
- Third, frequent promotions can condition customers to wait for discounts, reducing baseline demand in non-promotional weeks. To address this, SKUs can be grouped by promotional frequency and assigned a conditioning factor based on historical non-promotional demand relative to comparable low-promotion products.

These adjustments should remain planner-transparent and easy to override, while the resulting overrides and outcomes should be logged to improve future calibration.

.

5.3 Future Work

Several extensions could further improve the framework.
The most important opportunity is integrating additional causal drivers directly into the forecasting models. This includes structured data on promotions, pricing changes, trade events,

and customer-specific agreements. While Ramadan effects were incorporated through exogenous regressors, many other commercial drivers could not be modeled because clean historical data were unavailable.

A second extension is the introduction of cost-aware forecasting metrics. The current framework optimizes WMAPE, which treats all forecast errors equally. In practice, forecast errors have different business impacts depending on perishability, lead time, stockout risk, and holding cost. Incorporating these factors into model evaluation would better align forecasting decisions with operational and financial outcomes.

Finally, future work should evaluate ensemble approaches that combine multiple forecasting models within each cluster rather than selecting a single winner. Re-fitting the framework on the full active portfolio rather than the representative sample would also improve the robustness of the cluster-level accuracy estimates.

6. CONCLUSION

This capstone examined whether segmenting Isla Délice's SKU portfolio by demand behavior and assigning cluster-level forecasting models could improve demand-planning performance compared with a uniform forecasting approach.

Using K-means clustering on Average Demand Interval (ADI) and squared Coefficient of Variation (CV^2), the active portfolio was separated into four demand-pattern groups: Smooth, Intermittent, Erratic, and Lumpy. The results demonstrated that forecasting performance varies materially across these demand types and that no single model performs best across the full portfolio.

The analysis showed particularly strong results in the Smooth and Intermittent clusters, which together represent approximately two-thirds of active SKUs. AutoSARIMA performed best for Smooth demand patterns, while LightGBM performed best for Intermittent demand. ExpSmooth emerged as the most robust model for Erratic and Lumpy demand, particularly for its stability and reliability as a fallback in deployment.

Beyond model accuracy, the project demonstrates the operational value of segmentation itself. The cluster structure provides planners with a practical way to prioritize effort, distinguish between highly automatable and planner-intensive SKUs, and create a forecasting process that can evolve as demand behavior changes over time.

The broader contribution of this work is a transferable forecasting framework for food manufacturers and other FMCG companies managing large, heterogeneous SKU portfolios. The combination of ADI/CV^2 segmentation, cluster-level model assignment, and periodic cluster refreshes can be implemented without highly customized infrastructure and offers a scalable foundation for more advanced demand-planning capabilities.

REFERENCES

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. Holden-Day.

Boylan, J. E., & Syntetos, A. A. (2007). Accuracy and accuracy-implication metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, 4, 39–42.

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, 23(3), 289–303.

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, 38(4), 1583–1611.

Huang, T., Fildes, R., & Soopramanien, D. (2020). Forecasting retailer demand with machine learning and big data. *European Journal of Operational Research*, 284(2), 633–647.

Huiskonen, J. (2001). Maintenance spare parts logistics: Special characteristics and strategic choices. *International Journal of Production Economics*, 71(1–3), 125–133.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (pp. 281–297). University of California Press.

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.

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.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37–45.

Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, 214(3), 606–615.