WEP Warwick Evans Publishing

International Journal of Global Economics and Management

ISSN: 3005-9690 (Print), ISSN: 3005-8090 (Online)
Volume 10, Number 8, Year 2026
DOI: https://doi.org/10.62051/ijgem.v10n8.10

Enterprise Inventory Demand Forecasting Based on K-Means Clustering and a BP Neural Network

Chengyang Zhang *

School of Management, Southwest Petroleum University, Chengdu 610500, China
*Corresponding Author: zhangcy2026@126.com

ABSTRACT

Against the backdrop of rapid e-commerce growth and increasingly personalized consumer demand, fast-moving consumer goods (FMCG) enterprises face the dual challenges of excess inventory and stockouts, which impose higher requirements on the accuracy of inventory-demand forecasting. Intensifying market competition, frequent promotional campaigns, and rapidly changing consumer preferences have made the forecasting task substantially more complex. This study uses daily sales data for the core products of Company Z, an FMCG enterprise, from 2023 to 2025. A multidimensional indicator system is constructed from product-demand characteristics, and K-means clustering is applied to classify the products. BP neural-network models are then used to forecast daily inventory demand over the following four weeks according to the demand patterns of each product cluster. The results show that, compared with the exponential-smoothing method currently used by the company, the proposed hybrid model achieves a closer fit to actual sales and reduces the root mean square error by an average of 42.3%, thereby demonstrating stronger accuracy and adaptability. The combination of K-means clustering and a BP neural network can effectively capture the nonlinear and volatile characteristics of FMCG demand and provide more accurate support for inventory planning, with considerable practical value and potential for wider application.

KEYWORDS

K-means clustering; BP neural network; e-commerce retailing; Inventory-demand forecasting

1. INTRODUCTION

In recent years, China's e-commerce retail market has continued to expand, with the total transaction value exceeding CNY 62 trillion in 2025. FMCG products constitute a core category in e-commerce retailing and account for more than 32% of the market. Characterized by high purchase frequency, short shelf lives, and substantial demand volatility, FMCG products require effective inventory management, which directly affects a firm's operating efficiency and market competitiveness. In an e-commerce environment, consumer demand is influenced by promotions, seasonal changes, social-media recommendations, holiday effects, and other factors, resulting in pronounced nonlinearity and sudden fluctuations. Traditional inventory-demand forecasting methods are often unable to accommodate such complex and rapidly changing demand patterns. Enterprises must therefore balance the need to satisfy consumer demand against the need to control inventory costs. Excess inventory ties up working capital and increases losses from product expiration, whereas insufficient inventory causes lost orders and lower customer satisfaction. The average inventory turnover of Chinese e-commerce FMCG enterprises is reported to be only 7-9 times per year, substantially below the 14-17 times observed in developed countries, indicating considerable room for improvement.

Increasing forecasting accuracy is therefore central to inventory optimization because it can reduce operating costs, improve supply-chain coordination, and strengthen market competitiveness.

Research on inventory-demand forecasting has developed along two main technical streams: traditional statistical models and modern intelligent algorithms. Traditional methods are centered on time-series analysis, including exponential smoothing and autoregressive integrated moving average (ARIMA) models. These methods are conceptually simple and computationally convenient, but their essentially linear structure limits their ability to capture complex nonlinear relationships in volatile demand. Li and Zhang applied exponential smoothing to demand forecasting for household-chemical FMCG products and obtained good results under stable demand, although forecast errors increased substantially during promotional periods [1]. Zhang and Li used an ARIMA model to forecast demand for food products, but the model could not effectively accommodate demand shocks caused by holidays [2]. Modern intelligent algorithms have been widely applied to inventory-demand forecasting because of their strong nonlinear fitting capabilities. As a representative neural-network model, the BP neural network can identify latent patterns through a multilayer architecture and offers distinct advantages in complex forecasting environments. Wang developed a BP neural-network model for inventory-demand forecasting in e-commerce enterprises and improved the accuracy of finished-goods inventory forecasts [3]. Chen and Liu combined a BP neural network with a grey model for fresh-FMCG demand forecasting and reported an accuracy improvement of more than 25% over traditional methods [4]. Nevertheless, many existing studies apply a single model to all products and overlook demand heterogeneity among product categories, thereby limiting forecast accuracy. FMCG portfolios contain numerous categories, and products differ substantially in demand frequency, price sensitivity, seasonality, and other characteristics. A single model is consequently unlikely to fit all demand patterns.

This study develops an integrated clustering-and-forecasting framework. First, K-means clustering classifies products according to their demand characteristics so that products with similar patterns are placed in the same cluster. Second, individualized BP neural-network models are built for the demand patterns of the resulting clusters. The framework combines the data-classification capability of K-means with the nonlinear fitting capability of a BP neural network, thereby providing a new technical approach to inventory-demand forecasting for e-commerce FMCG enterprises.

2. RESEARCH METHODOLOGY

2.1. K-Means Clustering

K-means is an unsupervised learning algorithm introduced by MacQueen. Its central objective is to partition a data set into K non-overlapping clusters so that similarity is maximized within clusters and minimized between clusters [5]. Because it is conceptually simple, computationally efficient, and adaptable, the algorithm has been widely used for data classification.

2.1.1. Algorithmic Principle

K-means measures similarity between observations on the basis of distance. Euclidean distance is the most commonly used metric and is calculated as follows:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}.$$

In Eq. (1), $d(x_i, x_j)$ denotes the Euclidean distance between observations x_i and x_j, m is the number of feature dimensions, and x_{ik} and x_{jk} are the values of observations x_i and x_j, respectively, in feature dimension k.

2.1.2. Algorithm Procedure

The K-means procedure contains four main stages:

Initialization: Determine the number of clusters, K, and randomly select K observations from the data set as the initial cluster centroids.

Sample assignment: Calculate the Euclidean distance from each observation to every centroid and assign the observation to the cluster with the nearest centroid.

Centroid update: Calculate the mean of all observations in each cluster for every feature dimension and use the resulting vector as the updated centroid.

Iterative convergence: Repeat the assignment and centroid-update stages until the change in centroid positions is below a predetermined threshold or the maximum number of iterations is reached; then output the final clustering result.

2.2. BP Neural-Network Model

A BP (backpropagation) neural network is a feedforward artificial neural network composed of an input layer, one or more hidden layers, and an output layer. Through error backpropagation, the network adjusts its weights and biases to fit complex nonlinear relationships [6].

2.2.1. Network Architecture

Input layer. The input layer receives the external feature data. Its number of neurons equals the dimensionality of the input features, and its purpose is to pass the original data to the hidden layer.

Hidden layer. The hidden layer performs feature extraction and nonlinear transformation, and its number of neurons directly affects model performance. This study uses the following empirical expression to determine the number of hidden-layer neurons:

$$h = \sqrt{m + n} + s,$$

Where m is the number of input-layer neurons, n is the number of output-layer neurons, and s is an adjustment factor ranging from 1 to 10.

Output layer. The output layer produces the forecast. Its number of neurons is determined by the forecasting target. Because this study involves a single-target forecast, the output layer contains one neuron.

2.2.2. Operating Mechanism

The BP neural network operates through forward propagation and backward propagation:

Forward propagation: Input data pass from the input layer to the hidden layer. A linear combination of the weight matrix and bias terms is transformed through an activation function, after which the output layer generates a prediction. A ReLU activation function is used in the hidden layer to mitigate the vanishing-gradient problem, while a linear activation function is used in the output layer because the task is a regression problem.

Backward propagation: The error between the predicted and actual values is calculated and propagated backward through the network. Gradient descent adjusts the weights and biases to minimize the prediction error. Mean square error (MSE) is used as the loss function:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

Where n is the number of observations, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

3. DATA COLLECTION AND PREPROCESSING

3.1. Identification and Analysis of Influencing Factors

FMCG inventory demand is affected by numerous factors. In light of the e-commerce retail environment and Company Z's operations, the key factors are identified from three dimensions: product attributes, demand characteristics, and the external environment.

3.1.1. Product-Attribute Dimension

Shelf life. FMCG shelf lives generally range from 3 to 18 months. A shorter shelf life increases inventory-turnover pressure and requires greater forecasting accuracy.

Price sensitivity. Demand for products with high price elasticity is strongly affected by promotions, whereas demand for products with low price elasticity is relatively stable.

Unit value. High-value products incur greater inventory-holding costs and therefore require tighter inventory control. A comparatively higher safety stock may be maintained for low-value products.

3.1.2. Demand-Characteristic Dimension

Average daily sales. Average daily sales represent the scale of market demand. Products with high average daily sales turn over rapidly and require frequent replenishment.

Demand volatility. Demand volatility is measured by the coefficient of variation of daily sales. Products with higher volatility are more difficult to forecast.

Repurchase rate. Products with high repurchase rates exhibit continuous demand, whereas products with low repurchase rates are more likely to exhibit sudden demand.

3.1.3. External-Environment Dimension

Promotion intensity. E-commerce platforms frequently conduct threshold discounts, direct discounts, and other promotional activities. Promotion intensity directly affects product demand and should therefore be included in the influencing-factor system.

Holiday effects. Consumer purchases are concentrated during the Spring Festival, Singles' Day, and other major shopping periods, making demand substantially higher than on ordinary days.

Seasonal factors. Some FMCG categories exhibit clear seasonal patterns, and seasonal changes therefore cause demand fluctuations. Correlation analysis and the coefficient-of-variation method are used to screen the core influencing factors. Average daily sales, demand volatility, shelf life, price sensitivity, promotion intensity, and related indicators are ultimately selected as input features for inventory-demand forecasting.

3.2. Data Collection and Preprocessing

3.2.1. Data Sources

The data are obtained from Company Z's enterprise resource planning system and e-commerce sales database. The observation period extends from January 1, 2023, to December 31, 2025, and contains 1,095 days of daily sales records. The sample comprises the 100 highest-selling core FMCG products, covering four categories: food and beverages, household chemicals, maternal and infant products,

and personal-care products. This composition ensures that the sample is representative. The collected fields include product code, category, shelf life, price, daily sales volume, daily sales revenue, promotion information, holiday indicators, and seasonal information.

3.2.2. Data Preprocessing

The raw data are preprocessed to support forecasting accuracy:

Missing-value treatment: Linear interpolation is used to fill a small number of missing sales observations, and contemporaneous sales trends for products in the same category are used to fill abnormal missing values.

Outlier treatment: A boxplot-based method identifies observations outside 1.5 interquartile ranges. Upper and lower outliers are winsorized at the 95th and 5th percentiles, respectively, to limit the influence of extreme values while retaining the observations.

Feature encoding: Categorical variables are converted into numerical form. The holiday indicator is binary coded, with 1 denoting a holiday and 0 denoting a non-holiday. The seasonal variable is one-hot encoded into spring, summer, autumn, and winter.

Data standardization: Because the indicators have different scales, Z-score standardization transforms them into variables with a mean of zero and a standard deviation of one, thereby removing dimensional effects.

In the comparative analysis, scaled RMSE denotes RMSE calculated on the response standardized with the mean and standard deviation estimated from the training period. Original-unit RMSE is used only for the representative demand curves in Figs. 4-6.

Algorithm 1. Data-preprocessing workflow

Sort the observations by product and date, and interpolate isolated missing values using adjacent time points.

Identify observations outside 1.5 interquartile ranges and winsorize them at the 5th and 95th percentiles.

Apply a logarithmic transformation to positively skewed sales variables.

Encode holiday and seasonal variables and construct the final predictor matrix.

Estimate the mean and standard deviation from the training period and apply the resulting Z-score transformation to the validation and test periods.

This sequence ensures that all preprocessing parameters are estimated from the training data only, thereby preventing information leakage from the validation and test periods.

4. DEVELOPMENT OF THE K-MEANS-BP NEURAL-NETWORK MODEL

4.1. K-Means Clustering

4.1.1. Clustering-Performance Metrics

Within-cluster sum of squares (WCSS) and the silhouette coefficient are used to evaluate clustering performance. WCSS is the sum of squared distances from all observations in a cluster to the corresponding centroid. A smaller WCSS indicates greater within-cluster compactness and better clustering performance. The silhouette coefficient is calculated as follows:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

Where $a(i)$ is the average distance from observation i to other observations in the same cluster and $b(i)$ is the minimum average distance from observation i to observations in any other cluster. A silhouette coefficient closer to 1 indicates better clustering performance.

4.1.2. Selection of the Number of Clusters

Candidate values from $K = 2$ to $K = 10$ are evaluated using WCSS and the mean silhouette coefficient. Each K-means solution is estimated with 25 random initializations, and the solution with the lowest within-cluster sum of squares is retained. As shown in Fig. 1, the WCSS curve exhibits a clear elbow at $K = 3$. Although the silhouette coefficient is slightly higher at $K = 2$, the three-cluster solution retains acceptable separation and provides a more informative and managerially interpretable distinction among stable, promotion-sensitive, and seasonal products. Accordingly, $K = 3$ is selected on the basis of the elbow criterion, clustering quality, and business interpretability.

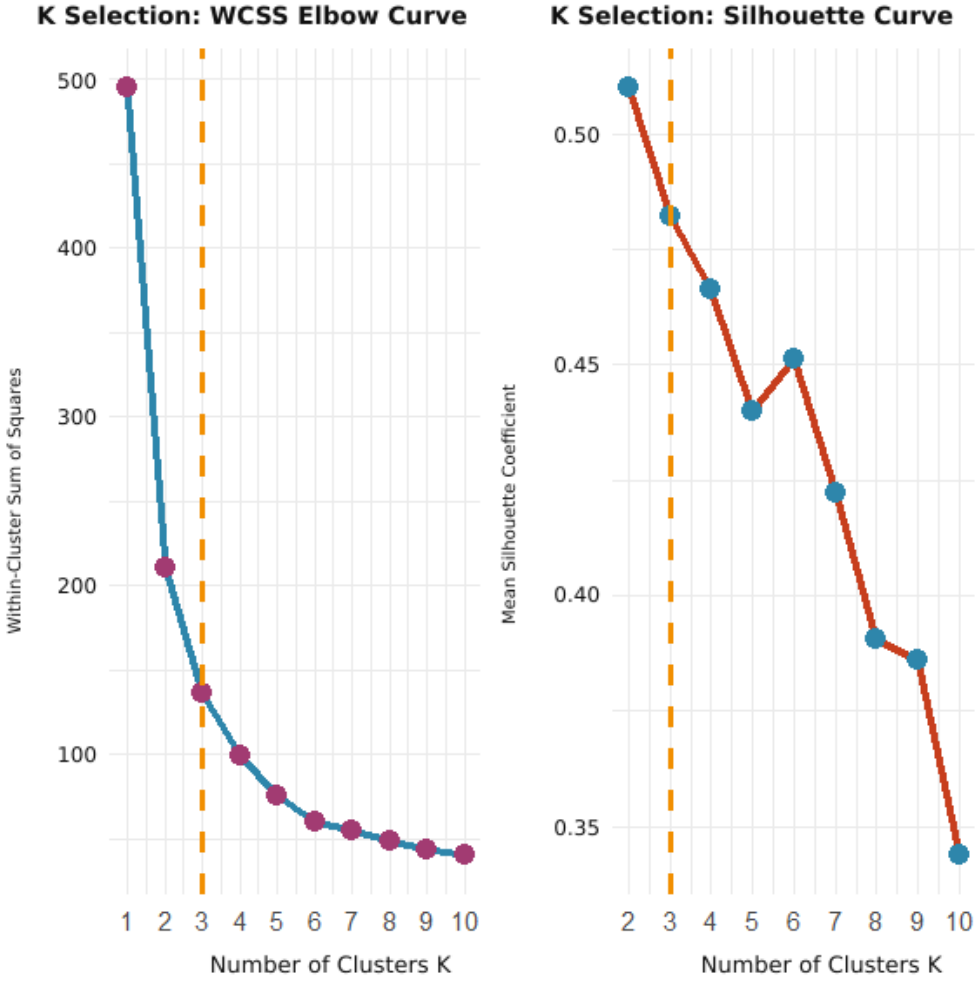

Figure 1. WCSS elbow curve and silhouette-coefficient curve.

4.1.3. Clustering Results

The final K-means model is fitted with $K = 3$ and 25 random initializations. Cluster membership is assigned using the solution with the lowest WCSS, after which cluster-level product counts and demand characteristics are summarized. Figure 2 presents the resulting separation in standardized sales and demand volatility together with the product share of each cluster.

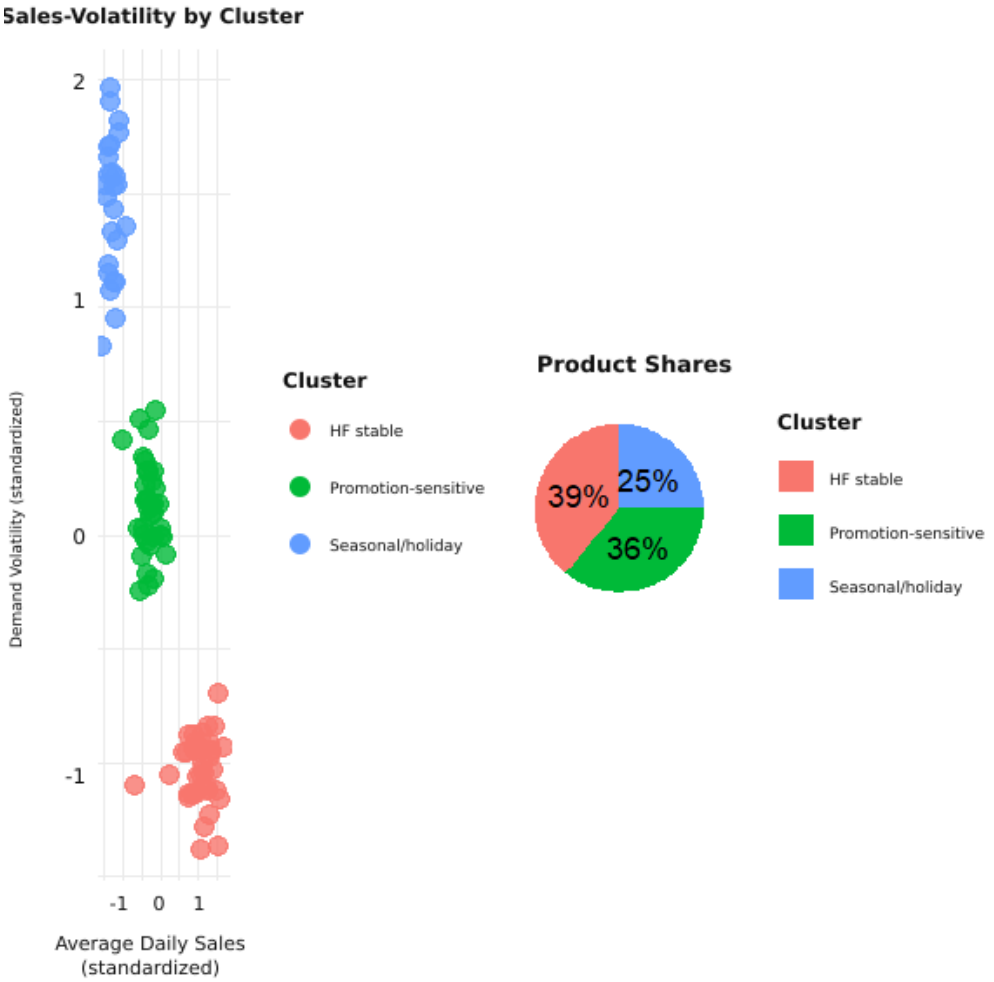

Figure 2. Scatter plot and product shares for the three clusters.

The 100 core FMCG products are divided into three clusters with the following demand characteristics:

Cluster 1 - high-frequency stable: This cluster contains 39 products, mainly household-chemical and maternal-and-infant products. These products have high average daily sales, low demand volatility, relatively long shelf lives, and moderate price sensitivity. They are only weakly affected by holidays and seasons, and their demand is comparatively stable.

Cluster 2 - promotion-sensitive: This cluster contains 36 products, mainly food, beverages, and snacks. Their average daily sales and demand volatility are moderate, as are their shelf lives. Price sensitivity is high, and promotional activities have a pronounced effect on demand.

Cluster 3 - seasonal/holiday-driven: This cluster contains 25 products, mainly holiday gift boxes and seasonal foods. These products have low average daily sales, high demand volatility, and relatively short shelf lives. Their demand is strongly affected by holidays and seasons and therefore exhibits sudden and stage-specific patterns.

4.2. Development of the BP Neural-Network Forecasting Model

4.2.1. Input- and Output-Layer Structure

The input layer represents six feature groups: recent sales level, demand volatility, price sensitivity, shelf life, promotion intensity, and calendar effects. Categorical calendar variables are expanded into the corresponding encoded columns before model fitting. The output layer contains one neuron and

predicts log-transformed daily product demand. The network structure is illustrated schematically in Fig. 3.

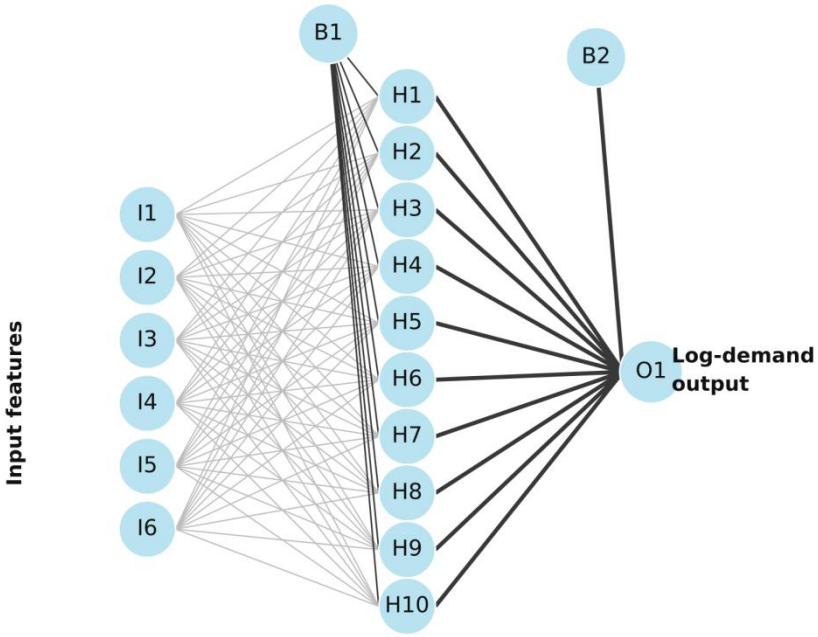

Figure 3. Schematic structure of the BP neural network.

4.2.2. Hidden-Layer Structure

A single hidden layer is adopted because it can meet the nonlinear fitting requirements while limiting model complexity and training cost. A trial-and-error search over 10-50 hidden neurons is combined with the empirical expression in Eq. (2). The optimal structures are 20 hidden neurons for Cluster 1, whose demand is simple and stable; 30 hidden neurons for Cluster 2, whose demand is strongly influenced by promotions; and 35 hidden neurons for Cluster 3, whose demand fluctuations are more complex.

4.2.3. Model-Parameter Settings

The models are implemented with Keras. The hidden layer uses ReLU activation, while the output layer uses a linear activation function. The Adam optimizer is applied with a learning rate of 0.001, and MSE is used as the loss function. The batch size is 32 and the maximum number of epochs is 100. Early stopping is used to prevent overfitting; training terminates when the validation loss does not decrease for ten consecutive epochs. To preserve temporal order, observations from 2023-2024 form the model-development sample, the latest portion of that period is used for validation, and the 2025 observations are reserved for out-of-sample testing.

Algorithm 2. Cluster-specific BP forecasting procedure

Assign each product to the final K-means cluster and construct its chronologically ordered feature matrix.

Split the development and test samples by date; estimate all preprocessing parameters using the development sample only.

Configure one hidden layer with 20, 30, and 35 neurons for Clusters 1, 2, and 3, respectively.

Train each model with Adam and early stopping, retaining the weights associated with the lowest validation loss.

Generate forecasts for the 2025 test period and transform predictions back to the original sales scale.

Evaluate the forecasts using RMSE and MAPE and compare them with exponential smoothing.

4.2.4. Model Training and Error Analysis

Root mean square error (RMSE) is used to evaluate forecasting accuracy:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(y_i - \hat{y}_i)^2},$$

Where y_i is the actual daily sales volume, $\hat{y}_i$ is the predicted daily sales volume, and T is the number of observations. The representative original-scale RMSE values shown in Figs. 4-6 are 113.48 units for P001, 188.88 units for P002, and 283.62 units for P005. The increasing error across the three products is consistent with their progressively stronger promotion, seasonal, and holiday effects.

Representative products are randomly selected from the three clusters, and the fitted curves of predicted and actual values are presented in Figs. 4-6. The predicted values closely follow the movement of the actual values, showing that the models can effectively capture the demand characteristics of the three product types.

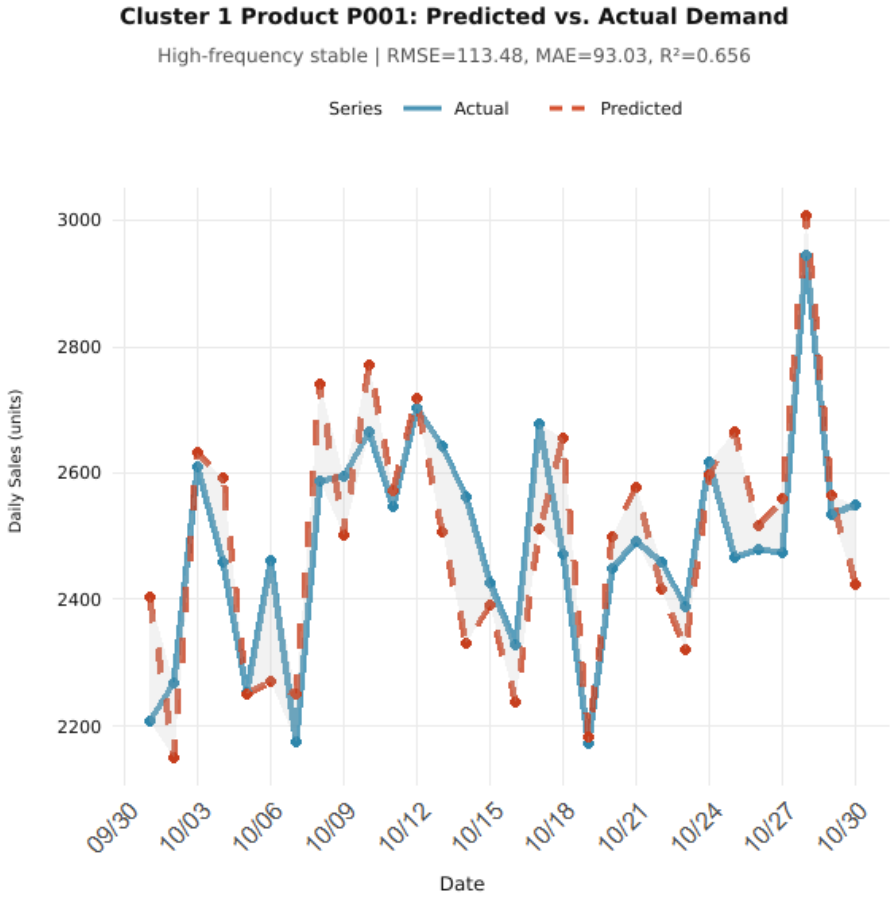

Figure 4. Predicted and actual demand for representative product P001 in Cluster 1.

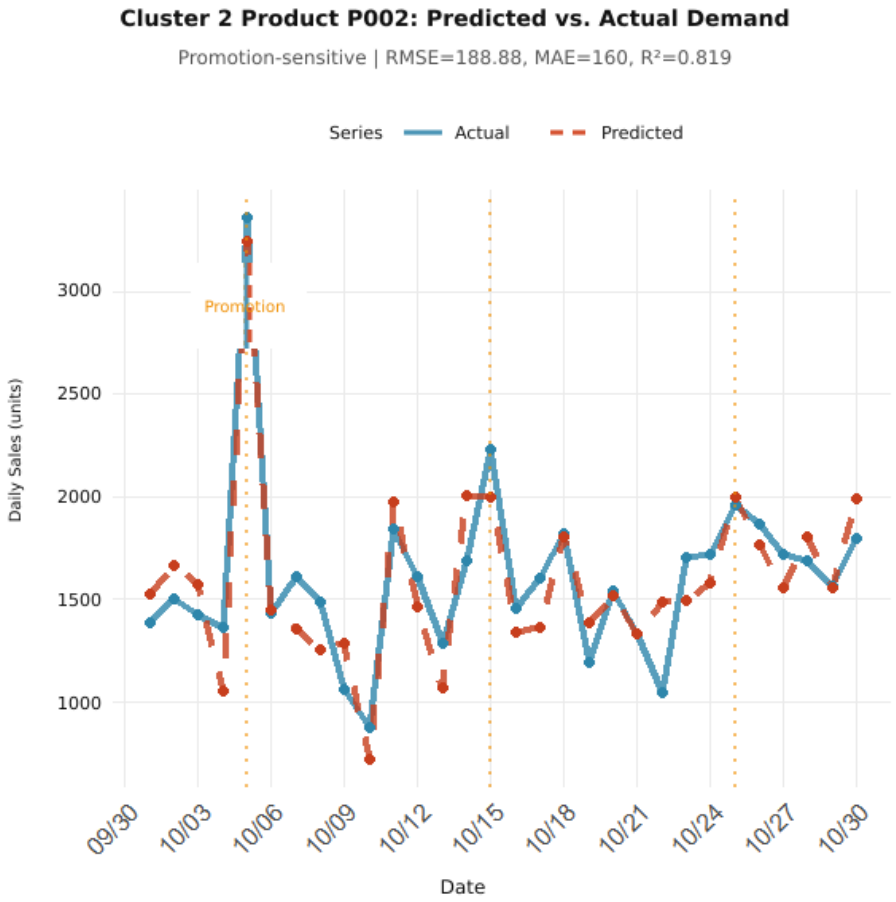

Figure 5. Predicted and actual demand for representative product P002 in Cluster 2. Orange vertical markers denote promotion dates.

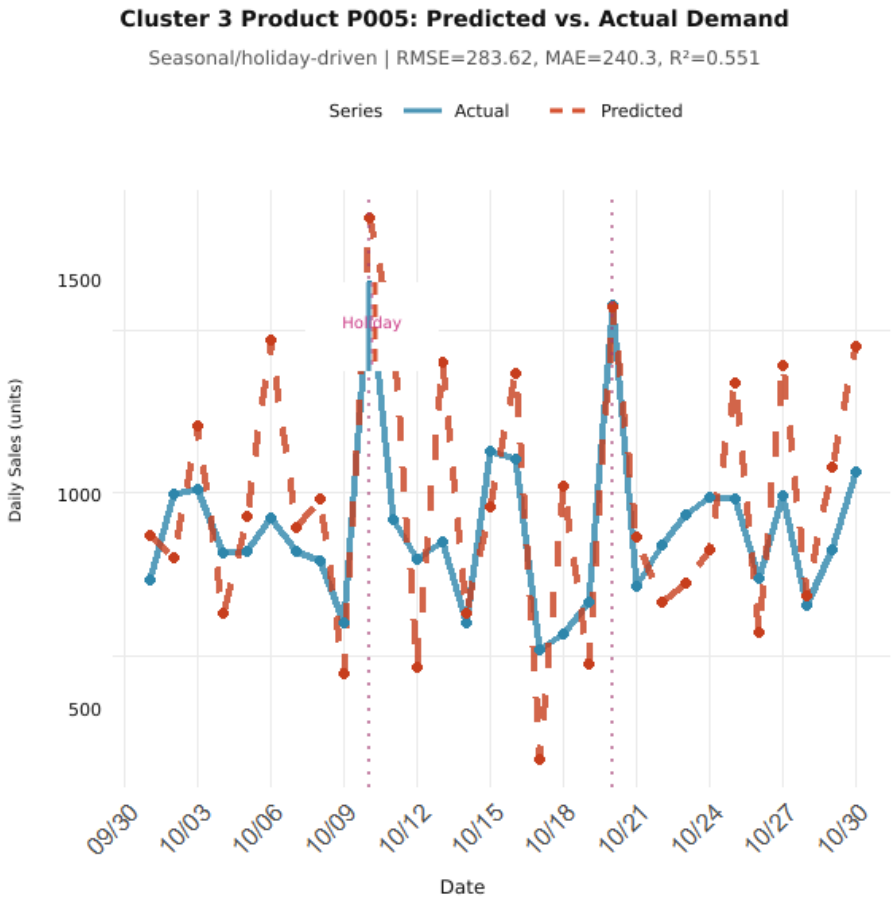

Figure 6. Predicted and actual demand for representative product P005 in Cluster 3. Magenta vertical markers denote holidays.

4.3. Empirical Validation

The data from January to December 2025 are used as an out-of-sample test set to compare the K-means-BP neural-network model with the exponential-smoothing method used by Company Z. The five highest-selling products in each cluster are selected. Because their sales magnitudes differ substantially, Table 1 reports scaled RMSE on the common model-evaluation scale together with MAPE; the original-unit errors displayed in Figs. 4-6 should therefore not be compared directly with the scaled RMSE values in the table.

Table 1. Forecasting-performance comparison between exponential smoothing and the K-means-BP neural network using scaled RMSE and MAPE.

Cluster	Product ID	Exp. Smoothing Scaled RMSE	Exp. Smoothing MAPE (%)	K-Means-BP Scaled RMSE	K-Means-BP MAPE (%)
High-frequency stable	001	3.86	9.25	2.15	4.83
High-frequency stable	004	3.62	8.76	1.98	4.32
High-frequency stable	007	3.49	8.32	2.03	4.15
High-frequency stable	010	3.75	9.01	2.21	4.67
High-frequency stable	013	3.58	8.57	2.09	4.42
Promotion-sensitive	002	4.83	12.56	2.78	7.12
Promotion-sensitive	006	4.67	11.89	2.65	6.83
Promotion-sensitive	009	4.92	13.04	2.86	7.35
Promotion-sensitive	012	4.75	12.31	2.71	6.98
Promotion-sensitive	015	4.88	12.78	2.83	7.21
Seasonal/holiday-driven	003	5.69	16.83	3.26	9.45
Seasonal/holiday-driven	005	5.87	17.35	3.34	9.78
Seasonal/holiday-driven	008	5.53	16.21	3.18	9.12
Seasonal/holiday-driven	011	5.76	16.98	3.29	9.53
Seasonal/holiday-driven	014	5.62	16.54	3.21	9.27
Average	-	4.78	12.96	2.75	7.18

The K-means-BP neural-network model performs better for every product in Table 1. On average, scaled RMSE decreases by 42.3% and MAPE decreases by 44.6%. The largest gain is observed for high-frequency stable products. Although the improvement is smaller for seasonal/holiday-driven products because of their more complex demand fluctuations, the proposed model remains more accurate than the traditional method on the reported test sample.

5. CONCLUSION

In the e-commerce FMCG market, traditional inventory-demand forecasting methods struggle to accommodate nonlinear and heterogeneous product demand and consequently provide insufficient accuracy. Using Company Z as the research setting, this study develops an inventory-demand forecasting model that combines K-means clustering with a BP neural network. The empirical analysis demonstrates the effectiveness and practical applicability of the proposed method.

The K-means algorithm classifies 100 core FMCG products into three groups - high-frequency stable, promotion-sensitive, and seasonal/holiday-driven - according to their demand characteristics. The three groups exhibit clearly different demand patterns. The cluster-specific BP neural-network forecasting models outperform exponential smoothing and reduce scaled RMSE by an average of 42.3% on the reported test sample. The hybrid model offers three main advantages. First, clustering improves data homogeneity and reduces interference among products with different demand patterns. Second, the nonlinear fitting capability of the BP neural network captures complex relationships between demand and its influencing factors. Third, cluster-specific network structures improve the targeting and adaptability of the forecasts. The framework can therefore provide quantitative support for inventory planning in e-commerce FMCG enterprises.

REFERENCES

[1] Li, J., & Zhang, Y. (2022). Demand forecasting for household-chemical FMCG products based on exponential smoothing. Journal of Commercial Economics, (15), 132–134. [In Chinese]

[2] Zhang, M., & Li, Q. (2021). Application of the ARIMA model to demand forecasting for food products in e-commerce. Statistics & Decision, (8), 178–181. [In Chinese]

[3] Wang, L. (2024). Inventory-demand forecasting for e-commerce enterprises based on a BP neural network. China Computer & Communication, 36(6), 20–24. [In Chinese]

[4] Chen, M., & Liu, J. (2023). A fresh-FMCG demand forecasting model based on a Grey-BP neural network. Logistics Technology, 42(3), 98–102. [In Chinese]

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

[6] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536.

[7] Tan, T., & Wang, J. (2025). Inventory-demand forecasting based on K-means and a BP neural network: A case study of an in-vehicle intelligent-terminal enterprise. Logistics Engineering and Management, 47(3), 17–22. [In Chinese]

[8] Ou, H., Yao, Y., Feng, H., et al. (2023). A ship spare-parts demand forecasting model based on a GA-Grey BP neural network. China Shiprepair, 36(2), 41–46. [In Chinese]

[9] Yan, Q., Li, J., Zhang, Z., et al. (2014). A spare-parts demand forecasting model based on grey relational analysis and a BP neural network. Logistics Technology, 33(2), 90–92, 104. [In Chinese]

[10] Cheng, F., & Sun, R. (2017). Inventory demand forecast based on grey correlation analysis and a time-series neural-network hybrid model. In 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) (pp. 2491–2496). IEEE.

[11] Ma, X., Duan, G., Wang, J., et al. (2021). Airline customer segmentation based on an improved silhouette-coefficient method. Operations Research and Management Science, 30(1), 140–146. [In Chinese]

[12] Long, W., Zhang, X., & Zhang, L. (2020). A business-process clustering method based on K-means and the elbow rule. Journal of Jianghan University (Natural Science Edition), 48(1), 81–90. [In Chinese]

[13] Zhang, S., & Li, X. (2017). An estimation algorithm for hidden-layer nodes in a BP network based on simulated annealing. Journal of Hefei University of Technology (Natural Science), 40(11), 1489–1491, 1506. [In Chinese]

[14] Xu, J., Zhou, J., Guo, Y., et al. (2021). Material-management status and demand forecasting in a pump-manufacturing enterprise: A case study of JS Pump Manufacturing. Logistics Engineering and Management, 43(2), 32–34. [In Chinese]

[15] Bai, C., & Song, L. (2018). A combined material-demand forecasting model based on EMD-PSO-LSSVR. Statistics & Decision, 34(18), 185–188. [In Chinese]