

Article | Received 20 February 2026; Revised 23 April 2026; Accepted 3 July 2026; Published 27 August 2026
<https://doi.org/10.55092/pcs20260007>

Short-form video consumption pattern analysis via hybrid grey relational K-means in big data scenarios

Tianshi Tan, Keren Mao, Xinyao Tan, Jiahao Zeng and Lisha Ma*

School of Computer Science, Hunan First Normal University, Changsha 410205, China

* Correspondence author; E-mail: mls2460169412@163.com.

Abstract: This study investigates the characteristics, influencing factors, and internal mechanisms of consumption preferences among short video e-commerce users. Analytical methods employed include grey relational analysis (GRA) and K-means clustering. Utilizing big data analytical techniques, this study identifies the key factors shaping user consumption preferences and constructs a detailed user profile framework. The GRA results indicate that monthly income, video browsing duration, and age significantly influence purchase frequency, with students' purchase frequency being most closely associated with occupational status. K-means clustering categorized users into five distinct groups with clear differences in consumption behavior and preferences. The findings contribute to provide both theoretical insights and practical guidance for precision marketing and operational optimization of short video e-commerce platforms.

Keywords: short-form video e-commerce; consumption preferences; grey relational analysis (GRA); K-means clustering; user segmentation.

1. Introduction

In recent years, short video e-commerce has rapidly become a key digital economy model, driven by content-based entertainment, social interaction, and precise algorithmic recommendations [1,2]. Mobile internet technologies and changing user behavior shape consumption decisions through creative content, immersive scenarios, and real-time interaction, with preferences influenced by content quality, social proof, and psychological immersion [3,4].

Empirical studies confirm the “content-social-scene” framework as key to preference reconstruction. Based on the SOR model, high-quality content (stimulus) evokes emotional resonance (organism), which then drives purchase intention(response) [1,2]. These mechanisms improve immersive experiences and align “user-context-temporal” factors, refining preference differentiation. Preferences vary by demographic: younger users prioritize entertainment and social connection, while older users value content utility [5,6]. Existing studies focus on isolated group analyses, lacking multidimensional frameworks for preference divergence. Technological advancements, like generative AI, enhance content distribution but also risk spreading misinformation [7,8], while supply chain integrity mechanisms emerge as critical for sustainable industry development.



Copyright©2026 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

Three research gaps persist: (1) overemphasis on behavioral outcomes neglects preference formation dynamics; (2) insufficient exploration of latent motivators; (3) inadequate integration of external variables. This study addresses these through a dual-method framework combining grey relational analysis (GRA) and K-means clustering.

GRA identifies core determinants (viewing duration, interaction frequency) through relational degree computation, while K-means stratifies users into behaviorally homogeneous clusters. Their synergy operates bidirectionally: GRA optimizes clustering variables to enhance accuracy, while clustered cohorts enable stratified GRA to reveal group-specific drivers. This integration systematically deciphers preference architectures, informing precision marketing strategies and advancing digital consumption theory.

2. Consumption preference analysis via grey relational theory and K-means clustering

This study applies grey relational analysis and K-means clustering to analyze short-video e-commerce consumer behavior [8]. Grey relational analysis identifies key factors influencing purchase frequency, while K-means clustering, after PCA dimensionality reduction, segments users and creates targeted profiles [9].

2.1. Mining key factors of consumer purchase behaviour based on grey relational analysis

This study collects data on age, gender, and occupation, using grey relational analysis to identify factors most impacting users' purchase frequency on short-video platforms.

To ensure the accuracy and effectiveness of the research, the data collected in this study were first encoded (for example, A was converted to 1, B to 2, and so on). Then, the encoded data were normalized using the Min-Max method. The calculation formula is as follows:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where X is the original data, X' is the normalized data, $X_{\min}$ and $X_{\max}$ are the minimum and maximum values in the dataset, respectively.

The grey relational coefficient $\gamma(x_0(k), x_i(k))$ measures the degree of correlation between the reference sequence $x_0(k)$ and the comparison sequence $x_i(k)$ at time k . The grey relational degree $\gamma(x_0, x_i)$ is the average value of the relational coefficients at all times, reflecting the overall degree of correlation between the sequences.

The grey relational coefficients of each core factor and the average monthly purchase frequency on the short-video platform are calculated as follows:

$$\gamma(x_0(k), x_i(k)) = \frac{\min \min |x_0(k) - x_i(k)| + \rho \max \max |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho \max \max |x_0(k) - x_i(k)|} \quad (2)$$

where $|x_0(k) - x_i(k)|$ is the absolute difference between the reference sequence and the comparison sequence at time k ; $\min \min |x_0(k) - x_i(k)|$ is the minimum absolute difference of all sequences at all times (i.e., the two-level minimum difference); $\max \max |x_0(k) - x_i(k)|$ is the maximum absolute difference of all sequences at all times (i.e., the two-level maximum difference); ρ is

the distinguishing coefficient, typically set to 0.5, with a value range of $[0, 1]$, used to adjust the sensitivity of the relational coefficient.

According to Equation (2), the grey relational degrees of each core factor and the average monthly consumption frequency on the short-video platform are calculated as follows:

$$\gamma(x_0, x_i) = \frac{1}{n} \sum_{k=1}^n \gamma(x_0(k), x_i(k)) \quad (3)$$

where n is the length of the sequence (*i.e.*, the total number of time points); $\gamma(x_0, x_i)$ is the relational coefficient at time k .

2.2. Construction of consumer portraits based on K-means clustering

To enhance the performance of K-means clustering, PCA is first applied to reduce dimensionality [10]. This preprocessing step simplifies the data structure and filters out noise, allowing K-means to operate more efficiently and effectively. By preserving key variance, PCA supports K-means in identifying underlying patterns, ultimately improving clustering accuracy and interpretability in high-dimensional settings.

2.2.1. Data dimensionality reduction via principal component analysis

To ensure Principal Component Analysis (PCA) accuracy and result reliability, this study standardized the original data to eliminate dimensional differences among features. The original data matrix is as follows:

$$X = (x_{ij})_{n \times p} \quad (4)$$

where n is the number of samples; p is the number of features; x_{ij} represents the j -th feature value of the i -th sample.

The calculation formula for the standardized feature value Z_{ij} is as follows:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad (5)$$

$$s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \quad (6)$$

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (7)$$

The standardized data matrix is denoted as $Z = (z_{ij})_{n \times p}$. After decomposing the features, principal components are selected. The contribution rate and cumulative contribution rate of the principal components are calculated. The calculation formulas are as follows:

$$r_k = \frac{\lambda_k}{\sum_{i=1}^p \lambda_i} \quad (8)$$

$$R_m = \frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^p \lambda_i} \quad (9)$$

where, r_k is variance proportion of k -th PC; λ_k is eigenvalue of k -th PC; λ_i is eigenvalue of i -th PC; and R_m is variance proportion of first m PCs.

2.2.2. Construction of the K-means clustering model for short-video user portraits

Randomly select k data points as the initial cluster centers, denoted as $c_1, c_2, c_3 \dots c_k$. Then, assign data points to the nearest clusters. For each data point x_i , calculate its distance $d(x_i, c_j)$, to each cluster center. Usually, the Euclidean distance is used, and the formula is as follows:

$$D(x_i, c_j) = \sqrt{\sum_{k=1}^d (x_{ik} - c_{jk})^2} \quad (10)$$

After each iteration, the center of each cluster needs to be recalculated. The calculation formula for the new cluster center c_j is as follows:

$$c_j = \frac{1}{|s_j|} \sum_{x_i \in s_j} x_i \quad (11)$$

where c_j is the coordinate of the j -th cluster center in the d -dimensional principal component space; $|s_j|$ represents the number of data points in the set s_j .

To ensure convergence, the algorithm stops iterating when the maximum shift of all cluster centroids between two consecutive iterations satisfies:

$$\max_i \|\mu_i^{(t)} - \mu_i^{(t-1)}\| < \varepsilon \quad (12)$$

where ε is a small threshold value (e.g., 10^{-5}) indicating convergence.

The K-means algorithm alternates between assigning data points to the nearest cluster center based on a distance metric and updating the cluster center, repeating this process until convergence or the preset iteration limit is reached.

Figure 1 illustrates the analytical framework for short-video e-commerce consumer behavior integrating grey relational analysis and K-means clustering, which takes data collection as the starting point and achieves multi-dimensional analysis through two parallel pathways:

(1) The grey relational analysis pathway: After coding and normalization processing, the analysis sequences are determined and the factor relational degrees are calculated to quantify the driving intensity of influencing factors on consumer behavior.

(2) The K-means clustering analysis pathway: Data quality is optimized via standardization and dimensionality reduction, and clustering is implemented iteratively to segment user groups, ultimately constructing behavioral portraits for each group.

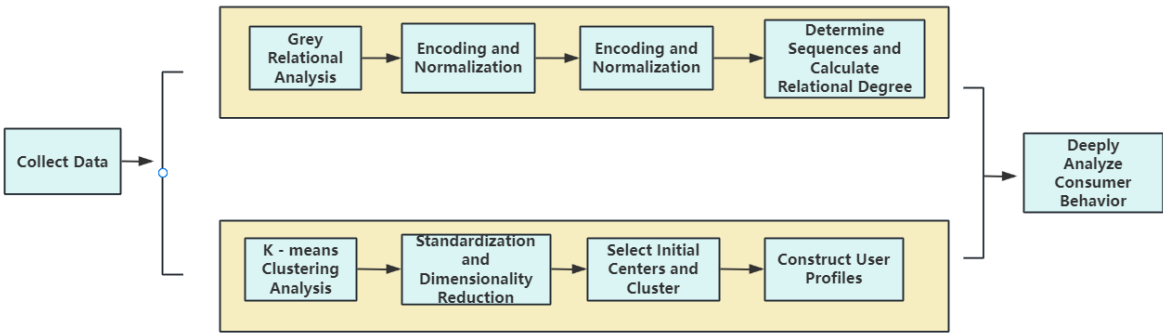


Figure 1. Analysis flow of consumer behavior based on GRA and K-means.

3. Model solution and result analysis

The experimental data of this study come from a questionnaire survey of residents in Changsha City, Hunan Province. Changsha is a new first-tier city in China, and studying the consumption preferences and habits of Changsha residents regarding short-video consumption is more representative.

The computer configuration used for our experiments is the 13th Gen Intel® Core™ i7–13620CPU (2.40 GHz) with Windows 11 64-bit. The programming software is PyCharm 2024.3, and the Python version is 3.12.

3.1. Results of grey relational analysis

The grey relational degrees of each core factor and average monthly purchase frequency via Equations (3) and (4) are shown in Table 1 and Figure 2. The grey relational degree of monthly income is highest at 0.9826, highlighting income as the core factor influencing purchase frequency. Video-watching duration follows at 0.9798, showing that longer viewing increases product exposure and promotes purchases. Age group scores 0.9790, indicating notable differences in purchasing behavior across age segments. Occupation (0.9412) and gender (0.9396) have relatively weaker impacts, with gender being the least influential among the analyzed factors.

Table 1. Grey relational degrees of key factors and purchase frequency.

Factor	Grey Relational Degree
Monthly Income Level	0.9826
Video-watching Duration	0.9798
Age Group	0.9790
Occupation	0.9412
Gender	0.9396

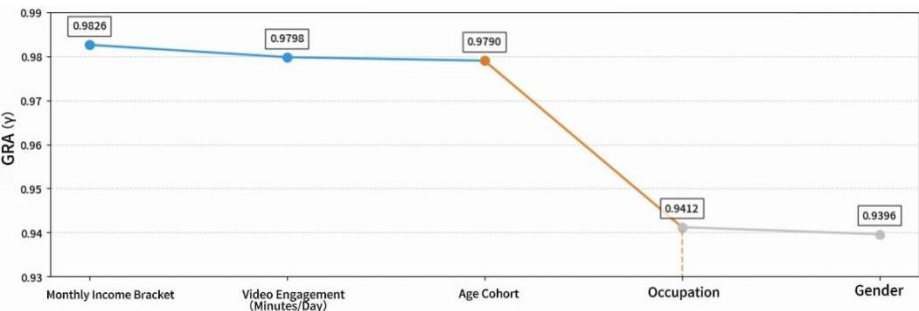


Figure 2. Grey correlation between factors and monthly purchases.

The grey relational degrees of each occupation and their average monthly purchase frequency via Equations (3) and (4) are presented in Table 2 and Figure 3. The grey relational value for students is highest at 0.9651, indicating a strong and regular correlation between student status and purchase frequency. Other occupations show slightly weaker but still significant correlations: “Others” at 0.9381, office workers at 0.9340, freelancers at 0.9286, and business owners at 0.9223, all suggesting occupation remains a relevant factor influencing purchase behavior.

Table 2. Occupation–purchase frequency grey correlation.

Occupation Factor	Grey Relational Degree
Student	0.9651
Others	0.9381
Office Worker	0.9340
Freelancer	0.9286
Business Owner	0.9223

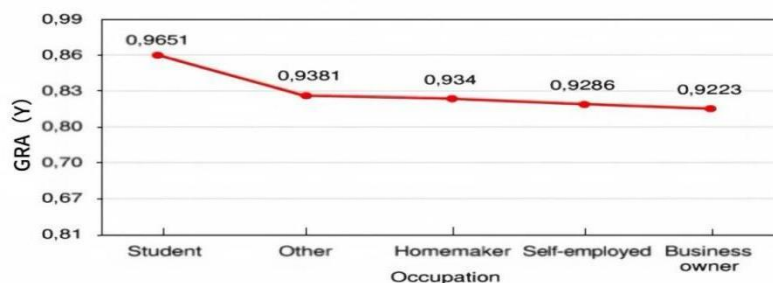


Figure 3. Grey correlation of occupation and monthly purchases.

3.2. Analysis of K-means clustering results

In this study, the number of principal components is determined by a cumulative contribution rate threshold of over 80%, ensuring they adequately capture the original data’s characteristics (higher rates indicate stronger explanatory power) [11]. Following principal component analysis, two principal components (named PC1 and PC2) with a 90% cumulative contribution rate were selected to retain core data variance, reduce dimensionality, and balance data reduction with information integration–supporting subsequent analysis [12]. The scatter plot of the K-means clustering results is shown in Figure 4.

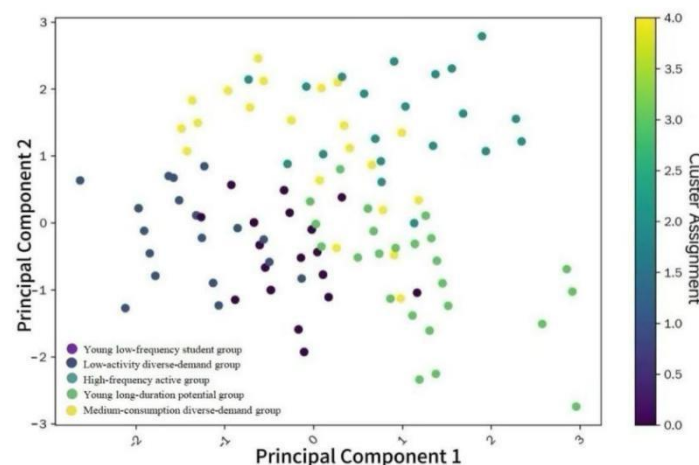


Figure 4. Scatter plot of K-means clustering analysis results.

After principal component calculation, two principal components with a cumulative contribution of 90% were selected in this study for the purpose of data dimensionality reduction. The 90% cumulative contribution rate is chosen as it balances variance retention and model simplicity. This common PCA threshold effectively captures most data variance while reducing dimensionality [11]. The two principal components were named as principal component 1 and principal component 2 respectively. The scatter plot of the *K*-means clustering results is shown in Figure 4.

Based on the results of *K*-means clustering analysis, short video platform users can be divided into five categories:

Based on *K*-means clustering analysis, users can be divided into five categories. Young low-frequency consumption students are mainly 18–25-year-old students, their consumption frequency is low and price-sensitive, the platform can push cost-effective commodities for their preferred categories such as apparel, beauty and makeup, and push promotional information such as school supplies during the school season and other nodes. Medium consumption diversified demand group covers 18–25 years old and 36–45 years old office workers, consumption needs are diverse, the platform needs to customize personalized recommendations in combination with their consumption habits, push high-end brand products and exclusive offers to enhance the conversion rate. The high-frequency consumption active group consists of students aged 18–25 and office workers aged 26–35, who have high consumption frequency and wide-ranging needs. The platform should accurately push new products, limited edition products and exclusive offers to satisfy their psychology of pursuing freshness and affordability. The young and long-time consumption potential group is composed of students aged 18–25, who brush short videos for a long time and have certain consumption potential. The platform can analyze their browsing and purchasing history, recommend niche or trendy commodities, and set up quality screening labels to promote purchasing. Low-activity multi-demand group includes people over 45 years old and under 18 years old, with low consumption activity. The platform needs to recommend similar high-quality products based on their purchase history, simplify the page for older users, and recommend trendy single products for younger users.

4. Conclusion

This study employs big data analytics, integrating grey relational analysis and the *K*-means clustering algorithm. Based on survey data from 163 participants in Changsha, it explores short-video e-commerce users' preferences. The research identifies monthly income, video-watching duration, and age as crucial factors affecting purchase frequency. Among occupations, the student status exhibits the strongest correlation with purchase frequency compared to others. Using the *K*-means clustering algorithm, the study segments users into five groups and constructs detailed user profiles, offering insights for consumer preference research and precision e-commerce strategies. Notably, due to the single-city data source, the user profile framework and research conclusions should not be over-generalized. Future research should incorporate multi-regional data, integrate diverse data sources like browsing behavior and social interactions with deep-learning algorithms, to verify and enhance the framework's universality.

Acknowledgments

The authors gratefully acknowledge the anonymous reviewers and editors for their insightful comments and constructive suggestions, which have significantly improved the quality of this manuscript. We also extend our sincere appreciation to all colleagues who provided assistance and support during this research. This work was supported by the Ministry of Education of the People's Republic of China under the Industry–University Cooperation Collaborative Education Project (Grant No. 250703879185935), and by the Enterprise Technology Research Project (Grant No. 2025–JSFW–047).

Authors' contribution

Conceptualization, methodology, software and writing—original draft, Tianshi Tan; formal analysis, software and visualization, Keren Mao; investigation and validation, Xinyao Tan; data curation and visualization, Jiahao Zeng; writing—review and editing, Lisha Ma. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

- [1] Chen X. Research on the impact of short–video marketing on consumer behavior based on SOR theory. *Mall Mod.* 2025, 4:20–22.
- [2] Kong K, Dong H. The effect of short–video marketing on consumers' impulsive purchase intention: The mediating role of online immersion and psychological intention. *Res. Bus. Econ.* 2024, 14:59–62.
- [3] Wang F, Liu X, Liu J, Zhang W, Chen C. How does scene–based empowerment drive stereoscopic value creation in short–video platforms? A case study of Racer. *Technol. Econ.* 2014, 43(6):44–57.
- [4] Li F. Risk prevention and control and countermeasures for “live streaming with goods” among specific groups. *Commer. Econ. Res.* 2021, 20:113–115.
- [5] She C. An empirical study on the consumption behavior of college students on the Jieyin platform. Master's thesis, Shanghai Normal University, 2024.
- [6] Cui R. Discussion on the problems and countermeasures of the live–e–commerce business model from a supply–chain perspective. *Commer. Econ. Res.* 2021, 8:69–71.
- [7] Zhang C. Development of the short–video industry in 2024: technological inflection point, cultural perspective, and commercial boiling point. *Mod. Audiovisual.* 2025, 1:29–34.
- [8] Li X, Wang Z. Combining grey relational analysis and clustering for consumer behavior analysis. *Int. J. Inf. Technol. Decis. Mak.* 2023, 22(3):891–916.
- [9] Li H, Zhang Y. Big data analytics for customer segmentation in E–commerce. *Electron. Commer. Res. Appl.* 2023, 56:101892.
- [10] Zhang Y, Wang F. User profiling in social media: a survey. *ACM Comput. Surv.* 2022, 55(1):1–38.
- [11] Wang BH. Multivariate statistical analysis and R language modeling, 5th ed. Beijing: Higher Education Press, 2024, pp. 224–225.
- [12] Jolliffe IT. Principal component analysis Springer–Verlag. *Int. Encycl. Stat. Sci.* 2025:1945–1948.