International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 6282-6292

https://doi.org/10.70517/ijhsa463539

A Cluster Analysis Study of Cross-border E-commerce Consumer Behavior under the Perspective of Digital Economy Globalization

Yingchao Lu[1,*] and Sijia Lv[2]
[1] School of Management, Seoul School of Integrated Sciences & Technologies (aSSIST University), Seodaemun, Seoul, 03600, Korea
[2] Harbin Medical University Cancer Hospital, Harbin Medical University, Harbin, Heilongjiang, 150081, China
Corresponding authors: (e-mail: yingchaolu1996@163.com).

Abstract The cross-border e-commerce market is booming in the context of digital economy globalization, and the accurate understanding of consumer behavior becomes the key to improve marketing efficiency. This study focuses on cross-border e-commerce consumer behavior characteristics under the perspective of digital economy globalization, and analyzes the shopping behavior data of 25,000 users provided by Tianchi Labs by using RFM customer segmentation model, entropy value method, factor analysis, and improved K-means clustering algorithm based on optimal K-value selection. The study extracted three public factors, namely, activity, purchase value and purchase intention, with a cumulative variance contribution rate of 83.45% through factor analysis, and constructed a cross-border e-commerce consumer behavior indicator system. The results of the cluster analysis show that cross-border e-commerce consumers can be divided into three categories: the first category of users (3079) with short consumption interval, high consumption frequency, low activity conversion rate, belonging to the important value customers; the second category of users (4209) with long consumption interval, low consumption frequency, low activity conversion rate, and low loyalty and satisfaction; the third category of users (17,712) with short consumption interval, low consumption frequency, low activity conversion rate, and remain active despite low loyalty. This study reveals the behavioral patterns and value differences of cross-border e-commerce consumers, providing data support and decision-making reference for platforms to develop differentiated marketing strategies.

Index Terms digital economy, cross-border e-commerce, consumer behavior, cluster analysis, K-means algorithm, factor analysis

I. Introduction

With the development of information and network technology, the globalization of digital economy has become a key point of global economic development [1], [2]. And the consumer group, as one of the economic subjects, will also be affected by cluster informatization and dataization [3]. In the face of the increasingly complex cross-border e-commerce environment, due to the different cultures, customs, and habits of each country, consumer behavior analysis has become an imminent topic, which is of great practical significance for understanding the similarities and differences of consumers in different countries, easing the dramatically expanding cross-border e-commerce business activities, and improving business synergy and cooperation among countries [4]-[7].

In the era of cross-border e-commerce, consumers' purchasing needs have changed considerably [8]. While traditional shopping in brick-and-mortar stores mainly focuses on physical goods, cross-border e-commerce provides richer choices [9], [10]. Consumers can easily purchase a variety of domestic and foreign brands through cross-border e-commerce platforms to meet personalized and diversified consumer demand [11], [12]. Consumer behavior, on the other hand, has an important impact on the development of cross-border e-commerce [13]. Consumer behavior on cross-border e-commerce platforms mainly includes browsing commodities, comparing prices, placing orders to buy and evaluating commodities, and is influenced by factors such as price, quality, service and trust, and is characterized by diversified choices, shopping experience, information transparency and social interaction [14]-[17]. By analyzing the consumer behavior of cross-border e-commerce, it is possible to provide high-quality products and services to meet the needs of consumers, establish a good brand image and reputation, and then promote the prosperity and development of the market [18]-[20].

The process of digital economic globalization has promoted the rapid development of cross-border e-commerce and profoundly changed consumers' shopping behavior and decision-making mode. Cross-border e-commerce has attracted a large number of consumers by virtue of its rich variety of commodities, obvious price advantage, convenient shopping experience and other characteristics, and the market scale has shown explosive growth. In

this context, accurately grasping the characteristics of consumer behavior and realizing effective classification and precise marketing of consumers have become the core elements for cross-border e-commerce platforms to enhance their competitiveness. Consumer behavior research, as the foundation of marketing, is of great significance for understanding the process of consumer decision-making, predicting purchasing behavior and formulating marketing strategies. In the past, research has mostly focused on consumer behavior in traditional retail contexts or single-country e-commerce platforms, and in-depth analysis of cross-border e-commerce consumer behavior is relatively insufficient. Especially in the era of big data, the behavioral trajectories of browsing, collecting, adding purchases, and purchasing left by consumers on the platform contain rich information on consumer preferences and decision-making patterns, which provides the possibility of in-depth excavation of consumer behavioral characteristics. Through the systematic analysis of these data, it is possible to comprehensively grasp consumers' purchasing habits, preferences, value contributions and potential needs, thus realizing consumer segmentation and value assessment. With the intensification of market competition and diversification of consumer needs, it is difficult to effectively differentiate consumer groups by traditional demographic or geographic variables. Consumer classification methods based on behavioral data can more directly reflect consumers' actual purchasing behavior and consumption patterns, providing more targeted guidance for the construction of user profiles, precision marketing and personalized recommendation of cross-border e-commerce platforms. In addition, an effective consumer segmentation strategy can also optimize platform resource allocation, improve marketing efficiency, enhance user experience and loyalty, and ultimately achieve a win-win situation for both platforms and consumers.

This study is based on the Taobao user shopping behavior dataset provided by Tianchi Labs, through the construction of cross-border e-commerce consumer behavior indicator system, applying RFM customer segmentation model to determine the key evaluation dimensions, introducing entropy method for objective weight assignment, using factor analysis to extract the public factors and optimize the indicator system, and finally applying improved K-means algorithm based on the optimization of K-value selection to realize consumer Cluster analysis. The study first preprocesses and standardizes the data to ensure the comparability of the data; second, extracts the key factors through factor analysis, reduces the dimensions and retains the main information; then, designs and implements the improved K-means clustering algorithm to improve the clustering effect by optimizing the selection of K-value; lastly, conducts an in-depth analysis and explanation of the clustering results to identify the behavioral characteristics and value contributions of different types of consumers.

II. Study design

With the acceleration of digital economy globalization, cross-border e-commerce has been developing rapidly in the Chinese market, attracting extensive attention from consumers. This paper will take cross-border e-commerce consumer behavior as the main body of research to explore the segmentation and clustering of cross-border e-commerce consumer behavior under the globalization of the digital economy, and improve the marketing efficiency and effectiveness of cross-border e-commerce.

II. A.Research data sources

User behavior analysis is conducted using the Taobao user shopping behavior dataset provided by Tianchi Labs. The dataset contains 2848225 non-repeated samples, recording the behavior of 25000 users over a period of time. The dataset contains five columns, which record the user ID, product ID, behavior category, product category, and operation time, and the behavior category contains four categories, which are browsing, bookmarking, adding shopping cart, and purchasing.

II. B.Pre-processing of research data

The research dataset contains cross-border e-commerce consumer historical behavior data. Preprocessing of the massive data,it is found that the dataset does not have missing data problem,so there is no need to fill in the missing values. Since each indicator describes different content and order of magnitude, the data of each indicator will be data standardized through data statistical transformation, using linear transformation to map the data to the same range in order to achieve comparability between the data. The corresponding data indicators finally obtained are shown in Table 1, covering the total number of operations, the number of products clicked by the user, the number of product types clicked by the user, the number of products added by the user, the number of product types added by the user, the interval of the most recent consumption, the frequency of consumption, and the amount of consumption.

Table 1: Cross-border e-commerce consumer behavior data indicators

Component	Index
1	Total operands
2	The number of items clicked by users
3	The number of product types clicked by users
4	Number of additional goods purchased by users
5	Number of types of goods purchased by users
6	Recent consumption time interval
7	Consumption frequency
8	Consumption amount

II. C.Research methodology

The research methods used in this research are RFM customer segmentation model, entropy method, factor analysis with improved K-means algorithm based on optimal K-value selection.

II. C. 1) RFM Customer Segmentation Models

The REM model is an important method used to measure the value of a customer and the customer's ability to generate profit, which measures the value of a customer by looking at when he/she recently purchased, how often he/she purchased, and how much he/she spent on the purchased goods [21]. The meaning of the three indicators used to construct the model and the results of specific customer segmentation) are as follows:

1) Last purchase: refers to the time of the customer's last purchase. Theoretically, the more recent the last purchase time of the customer, the easier to grasp the customer. Of course this data is always changing, every time a customer makes a purchase it will change once. Currently, according to research, the smaller the R-value, the shorter the time interval since the last head purchase, and the greater the chance that the customer will buy the product at one time. Similarly, the smaller the R, the more complete the company retains the customer's information, while the larger the R, the less the company retains the customer's information, and if the time interval is more than half a year, the company's information on the customer is 50% invalid.

2) Consumption frequency: the number of times a customer consumes at a certain time. The greater the consumption frequency, the greater the customer's loyalty to the company, the stronger the dependence on the company. The most frequently purchased customers are often the most loyal customers, to increase the number of times this part of the customer's purchases, which means that from the competitors to seize the market share.

3) Consumption amount: the amount of money spent by the customer to buy goods in a certain period of time, which is the most important indicator, and also verifies the “Pareto's Law”, 80% of the company's revenue comes from 20% of the customers. The greater the amount of money spent, the greater the contribution made by the customer to the company.

II. C. 2) Entropy method

The classical method of objective assignment method is the entropy method [22]. This method is designed to reduce the adverse effects caused by subjective factors and to utilize data to solve objectively. The most important feature of the entropy value method is that it directly uses the actual information to calculate the weights without adding subjective judgments. The specific steps are as follows:

1) Assume a multi-attribute decision matrix (1), each row of the matrix represents a program a_i and each column represents an attribute a_j:

$$Q = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \tag{1}$$

2) Calculate the contribution of the i th program a_i under the j th attribute using equation (2):

$$P_{ij} = a_{ij} / \sum_{i=1}^{m} a_{ij} \tag{2}$$

3) Calculate the total contribution of all programs to attribute a_j with equation (3) as follows:

$$E_j = -K\sum_{i=1}^{m} P_{ij}\ln(P_{ij}) \tag{3}$$

where K equation (4) is given below:

$$K = 1/\ln(m) \tag{4}$$

This ensures that the total contribution E_j is between [0, 1].

4) Use Eq. (5) to find out the consistency of the contribution of each program under the attribute j:

$$d_j = 1 - E_j \tag{5}$$

5) To find the weights, equation (6) is as follows:

$$W_j = d_j / \sum_{j=1}^{m} d_j \tag{6}$$

When $d = 0$, the attribute j can be deleted because its weight is equal to 0. If the subjective estimation of the weights λ_i is already available beforehand, the weighting formula can be corrected as formula (7):

$$W_j = \lambda_j d_j / \sum_{j=1}^{m} \lambda_j d_j \tag{7}$$

II. C. 3) Factor analysis

Factor analysis is a multivariate simplification technique that aims to decompose the original variables and find potential categories from the original variables, placing variables with strong correlations between indicators in one category and lower correlations between different categories. Each category of variables represents a “common factor” and these few “common factors” are then used to solve the problem instead of the original indicators. The factor model is shown in equation (8):

$$\begin{aligned} X_1 &= f_{11}g_1 + f_{12}g_2 + \ldots\ldots + f_{lm}g_m + \alpha_1 \\ &\vdots \\ X_n &= f_{nl}g_1 + f_{n2}g_2 + \ldots\ldots + f_{mn}g_m + \alpha_n \end{aligned} \tag{8}$$

where the number of samples is set to be T, n indicators, $X = (x_1 \ldots x_n)$ is the random vector, the common factor to be found is $f = (f_1 \ldots f_m)^T$, and the matrix $g = (g_{ij})$ is known as the factor loading matrix, g_{ij} is the factor loadings, and the factor loadings a_{ij} refer to the loadings of the variables at the i th variable on the i th factor, reflecting the importance of the i th variable on the i th factor. The α represents influences other than those of the common factors, but are generally ignored.

For the derived common factor, it is necessary to find those indicators on which its loadings are higher and then interpret the common factor according to that. However, when there is no way to give an explanation, it is necessary to make further factor rotation in the hope that a more reasonable explanation can be obtained through the redistribution of factor loadings. Two features of the model: first, the model is not affected by the magnitude: second, the factor loadings can get a new factor loading matrix through factor rotation, which makes the interpretation of the common factor more reasonable. But factor analysis has some caveats:

1) KMO test. The KMO test is used to test the partial correlation between variables, taking a value between 0 and 1. Theoretically, the closer the KMO statistic is to 1, the stronger the bias correlation between the variables is, and the better the analysis will be. In practical application, the effect will be better if the KMO statistic is above 0.7; when the KMO statistic is under 0.5, factor analysis is not applicable.

2) Each common factor in factor analysis has practical significance, while each principal component in principal component analysis is so important. However, in factor analysis, if the extracted common factor has no practical significance, it is necessary to re-measure each index.

II. C. 4) Improved K-means algorithm based on optimal K-value selection

1) Description of K-means algorithm

K-means is a classical division-based clustering method, which generally uses the Euclidean distance as a measure of the similarity between two data points, the greater the similarity, the smaller the distance [23]. The core idea of the algorithm is: first determine the number of clusters K and K initial clustering centers. According to the distance between the data points and the clustering centers, the position of the clustering centers is constantly updated, making the sum of squared errors (SSE) of each cluster smaller. When the SSE no longer changes or the objective function converges, the SSE value is minimized and the iteration stops to obtain the final clustering result. The algorithm flow is as follows.

(1) Initialize the clustering center

Determine the number of clusters K. Randomly select K points from the dataset as the initial clustering center $C_i(1 \le i \le K)$.

(2) Assigning samples

Calculate the Euclidean distance between the remaining data points and the clustering center C_i, find the shortest distance and assign all samples to the cluster corresponding to the clustering center C_i.

The Euclidean distance is calculated as:

$$d(x, C_i) = \sqrt{\sum_{j=1}^{m} (x_j - C_{ij})^2} \tag{9}$$

In Eq. (9), x is the data object, C_i is the i th clustering center, m is the dimension of the data object, and x_j, C_{ij} is the j th attribute value of x and C_i.

(3) Update the clustering center Calculate the mean and squared error of all points in each cluster, use the mean as the new clustering center and repeat step (2).

The squared error is calculated as

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} | d(x, C_i) |^2 \tag{10}$$

(4) Until the clustering center no longer changes or reaches the maximum number of iterations, the cycle ends and the final clustering results are obtained.

2) Improved K-means algorithm for optimizing K value selection

Although the K-means clustering algorithm does not rely on the label information of the data and has a good interpretation effect on big data feature mining, the traditional K-means algorithm has the defect of artificially setting the K value. Based on this, this paper introduces the Calinski-Harabasz (CH) clustering quality evaluation index, and sets the class corresponding to the highest CH score as the number of clusters in this study.

The CH metric is obtained from the ratio of the inter-cluster sample separation to the intra-cluster sample compactness, with a larger CH representing the more compact the class itself and the more dispersed the classes are, i.e., a better clustering result. When the intra-cluster is dense and the inter-cluster separation is better, the optimal number of clusters can be clearly derived from the CH score line graph, and it has the advantage of fast calculation speed.

The implementation steps of the K-means algorithm for CH metric optimization are as follows. Input: dataset $X = \{x_1, x_2, \cdots, x_n\}$ (n denotes the number of data points).

(1) Determine the optimal number of clusters K.

In the first step, WGSS is calculated.

WGSS (Within-Groups Sum of Squared Error) is the Within-Groups Sum of Squared Error, which is used to measure the tightness of the samples within the clusters, the smaller the WGSS, the tighter the clusters are, and the better the clustering effect. The formula is

$$WGSS = \frac{1}{2} \left[(n_1 - 1)\bar{d}_1^2 | \cdots | (n_k - 1)\bar{d}_k^2 \right] \tag{11}$$

In Eq. (11), $\bar{d}_k^2$ is the average distance of the samples within the k th cluster and n_k is the number of samples in the k th cluster.

In the second step, the BGSS is computed.

BGSS is the sum of squared errors between clusters, which is used to measure the separation of samples between clusters; the larger the BGSS is, the more dispersed the clusters are, and the better the clustering effect is.

Its calculation formula is

$$BGSS = \frac{1}{2}\left[(k-1)\bar{d}^2 + \sum_{i=1}^{k}(n_i - 1)(\bar{d}^2 - \bar{d}_i^2)\right] \tag{12}$$

In Eq. (12), $\bar{d}^2$ is the average distance between all the samples, $\bar{d}_i^2$ is the average distance of the samples within the cluster of the i th cluster; n_i is the number of samples in the i th cluster; and k is the number of clusters in the sample set.

In the third step, the CH value is calculated.

The mathematical formula for the CH score value S is:

$$\begin{aligned} S(k) &= \frac{\text{The sum of squared errors between clusters}}{k-1} \\ &\quad / \frac{\text{The sum of squared errors within the cluster}}{n-k} \\ &= \frac{BGSS}{WGSS} \times \frac{n-k}{k-1} \end{aligned} \tag{13}$$

The smaller the WGSS and the larger the BGSS, the larger the value of the CH indicator and the better the clustering effect.

In the fourth step, the CH line graph is plotted to determine the optimal number of clusters K.

(2) Run K-means to derive clustering results:

In the first step, initialize the clustering center.

Input the number of clusters k, and randomly select k points from the data set as the initial clustering center $C_i(1 \le i \le K)$.

In the second step, assign samples.

Calculate the Euclidean distance between the remaining data points and the clustering center C_i, find the shortest distance and assign all samples to the cluster corresponding to the clustering center C_i.

In the third step, update the clustering center.

Calculate the mean and squared error of all points in each cluster. Use the average value as the new clustering center and repeat the second step.

Step 4, until the clustering center no longer changes or the maximum number of iterations is reached, the loop ends and the final clustering result is obtained.

In the fifth step, output the clustering result: $C = \{c_1, c_2, \cdots, c_k\}$.

III. Factor analysis of cross-border e-commerce consumer behavior evaluation indicators

III. A. Factor analysis applicability test

Before using factor analysis methods, it is necessary to test the suitability of the research data for factor analysis. In this paper, KMO and Bartlett's test of sphericity are used to test the suitability of factor analysis. Among them, the larger the KMO value, the stronger the correlation between the variables, the more suitable for factor analysis. If the KMO value is less than 0.5, it proves unsuitable for factor analysis. And when the significance of the chi-square statistical value of Bartlett's test of sphericity is less than 0.05, it can be considered that there is a significant difference between the correlation coefficient matrix and the unit matrix. Using SPSS25 statistical software, the results of factor analysis KMO and Bartlett's sphericity test were obtained as shown in Table 2. As can be seen from the table, the KMO value of 0.688>0.5 was obtained and the chi-square statistical significance of Bartlett's test of sphericity was 0.001<0.05, indicating that the research data is suitable for factor analysis.

Table 2: KMO and Bartlett's sphericity test results

Sampling appropriateness quantity		0.688
Bartlett 's sphericity test	Approximate chi-square	9242.45
	Degree of freedom	30
	Prominence	0.001

III. B. Extraction and Determination of Common Factors

Based on the processed data, it was analyzed by principal component analysis using SPSS25 software to extract the public factors with eigenvalues greater than 1, and finally the variance interpretation of each component was obtained as shown in Table 3. As can be seen from the data in the table, the cumulative variance contribution rate of the three public factors extracted from the eight variables through factor analysis is 83.45%, indicating that the three extracted public factors carry 83.45% of the information of the previous eight original variables, which can explain most of the cross-border e-commerce consumer behavior evaluation indicators.

Table 3: Total variance explanation

Component	Initial eigenvalue			Extracting the sum of squared loads			The sum of square of rotating load		
	Total	Percentage variance	Accumulation (%)	Total	Percentage variance	Accumulation (%)	Total	Percentage variance	Accumulation (%)
1	3.395	42.4375	42.4375	3.395	42.4375	42.4375	2.45	30.625	30.625
2	2.291	28.6375	71.075	2.291	28.6375	71.075	2.151	26.8875	57.5125
3	0.99	12.375	83.45	0.99	12.375	83.45	2.075	25.9375	83.45
4	0.417	5.2125	88.6625	-	-	-	-	-	-
5	0.433	5.4125	94.075	-	-	-	-	-	-
6	0.208	2.6	96.675	-	-	-	-	-	-
7	0.201	2.5125	99.1875	-	-	-	-	-	-
8	0.065	0.8125	100	-	-	-	-	-	-

The fragmentation diagram of each component is shown in Figure 1. From the figure, it can be seen that the eigenvalues of the first three common factors are relatively large, which are all larger than 1. Until after the fourth common factor, the eigenvalues gradually tend to stabilize with little change. Combined with the data of the fragmentation diagram, it is more certain that the three extracted eigenfactors are reasonable.

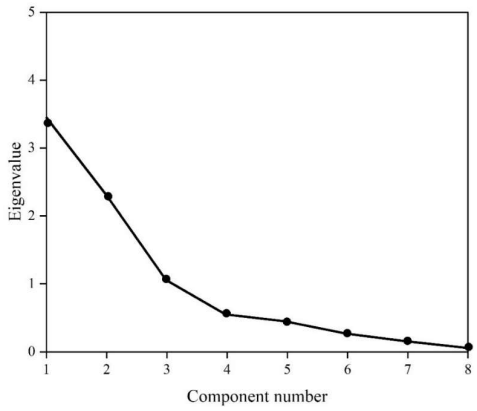


Figure 1: Stone map

III. C. Rotation and naming of common factors

In this study, SPSS25 statistical software was used to perform orthogonal rotation of each component by Caesarized maximum variance method in the process of factor analysis of the research data, and the rotated component matrix was obtained, as shown in Table 4. According to the data in the above table, a total of three common factors were extracted through factor analysis of cross-border e-commerce consumer behavior evaluation index data. In the

factor molecule, each common factor is a linear combination of various variables, if a variable has a large load component on a common factor, it means that the common factor can better reflect the information of the variable. The three common factors are named "activity", "purchase value", and "purchase intention".

Table 4: Rotated component matrix

-	Component		
	1	2	3
Total operands	0.905	0.087	0.196
The number of items clicked by users	0.976	0.023	0.046
The number of product types clicked by users	0.87	0.001	-0.04
Number of additional goods purchased by users	0.104	0.281	0.902
Number of types of goods purchased by users	0.077	0.194	0.944
Recent consumption time interval	0.107	0.637	0.465
Consumption frequency	0.108	0.891	0.307
Consumption amount	-0.026	0.91	0.064

III. D. Cross-border e-commerce consumer behavior indicators identified

Through the above factor analysis, three public factors are extracted, and the final e-commerce customer value evaluation index system is shown in Table 5. According to the data in the table, the activity index and the purchase intention index are indicators that can indicate the customer demand structure, while the potential consumption behavior of customers can be measured using the customer demand structure. Overall, it can be said that the results of the factor analysis and the established cross-border e-commerce consumer behavior indicator system are largely compatible, which verifies the rationality of the constructed cross-border e-commerce consumer behavior indicator system.

Table 5: Cross-border e-commerce consumer behavior index system

Goal	Common factor	Indicators	Correlation coefficient
Cross-border e-commerce consumer behavior	Purchase value	Recent consumption time interval	0.627
		Consumption frequency	0.862
		Consumption amount	0.902
	Activity	Total operands	0.913
		The number of items clicked by users	0.959
		The number of product types clicked by users	0.845
	Purchase intention	Number of additional goods purchased by users	0.895
		Number of types of goods purchased by users	0.959

IV. Cross-border e-commerce consumer behavior cluster analysis

In this chapter, the improved K-means algorithm based on optimized K-value selection proposed in this paper will be further adopted for cross-border e-commerce consumer behavior category classification.

IV. A. Optimal K determination

Firstly, the K value is determined by using the elbow method and the contour coefficient, and the results of recalculation using the elbow method are shown in Fig. 2. The results show that the optimal K value is 3 or 4, and the contour coefficients are calculated respectively, the contour coefficient when K is taken as 3 is 0.576,,and the contour coefficient when K is taken as 4 is 0.492, so the final determined optimal K value is 3.

IV. B. Analysis of clustering results

The cross-border e-commerce consumers are classified into three categories by the improved K-means algorithm based on optimal K-value selection, as shown in Fig. 3. 25,000 cross-border e-commerce consumers are classified into three categories, among which the number of users with category label 2 is the largest, 17,712; the number of users with category label 1 is the second largest, 4,209; and the number of users with category label 0 is the smallest, 3,079.

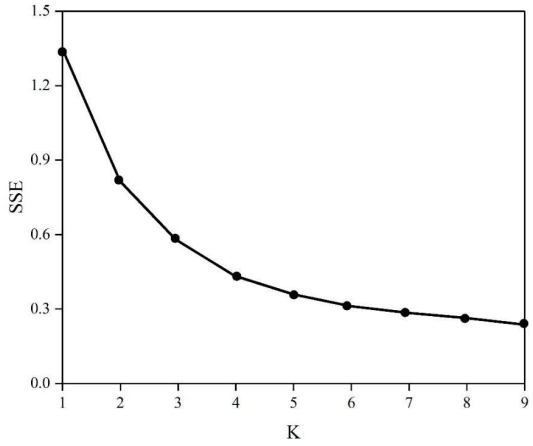


Figure 2: Calculation result diagram of elbow method

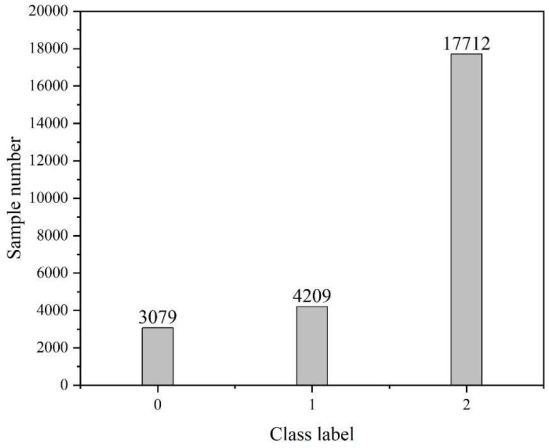


Figure 3: Clustering result diagram

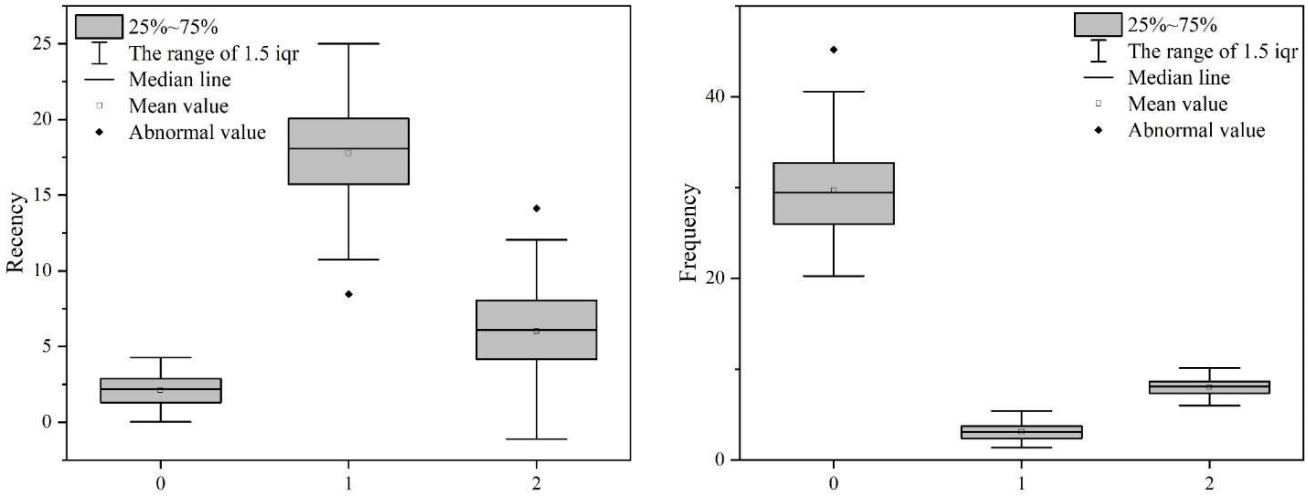
For each type of user, box line graphs are plotted, as shown in Figure 4, with graphs (a) and (b) corresponding to the consumption practice interval, and consumption frequency, respectively. The graphical results show that:

1) The cross-border e-commerce users with label 0 have a lower consumption interval, which indicates that there is still shopping behavior in the near future, and their consumption frequency is higher, but the conversion rate of activity is lower, which on the one hand, indicates that this type of user has higher loyalty and satisfaction, and on the other hand, indicates that they are more active, and they are able to generate higher value.

2) The cross-border e-commerce users with label 1 have a higher consumption time interval, indicating that they have not generated shopping behavior on the cross-border e-commerce platform for a longer period of time, and their consumption frequency is lower and their activity conversion rate is also lower, which on the one hand indicates that the loyalty and satisfaction of this type of users are lower, and on the other hand, it also indicates that they still maintain a certain degree of activity.

3) The cross-border e-commerce users with label 2 have a lower consumption interval, which indicates that there is still shopping behavior recently, and their consumption frequency is lower and the conversion rate of activity is also lower, which on the one hand indicates that the user's loyalty and satisfaction are lower, and on the other hand also indicates that he or she still maintains a certain degree of activity.

In summary, the cross-border e-commerce users with label 0 belong to the important value customers, and there are 3079 users in this category, which is in line with the “law of two or eight”, i.e., 80% of the profits are often created by 20% of the consumers, further confirming the validity and reliability of the clustering results.



(a)Recency (b)Frequency

Figure 4 Experimental results box plot

V. Conclusion

By analyzing the behavioral data of 25,000 cross-border e-commerce users, the study clearly delineates three types of consumer groups with different characteristics, which provides a theoretical basis for the precision marketing of cross-border e-commerce platforms. The results of factor analysis show that the three public factors of activity, purchase value and purchase intention can explain 83.45% of the variation in consumer behavior, constituting a core indicator system for evaluating cross-border e-commerce consumer behavior. Cluster analysis found that the first category of users accounted for 12.31%, showing the characteristics of short consumption interval, high consumption frequency and low conversion rate of activity, which are the core value customers of the platform; the second category of users accounted for 16.84%, with a long consumption interval, low consumption frequency and low loyalty and satisfaction, but still maintain a certain degree of activity; the third category of users has the largest number of users, amounting to 17,712, accounting for an overall of The third category has the largest number of users, amounting to 17,712, accounting for 70.85% of the total. Although the consumption frequency is not high, the consumption time interval is short, and the activeness of the platform is maintained. This distribution pattern verifies the applicability of the “law of two or eight” in cross-border e-commerce, i.e. about 12% of core users create the main value of the platform.

Based on the clustering results, cross-border e-commerce platforms should adopt differentiated marketing strategies for different types of consumers: implement exclusive member services and loyalty incentive programs for the first type of core value users; carry out recall activities and personalized recommendations for the second type of users; and provide richer product experiences and interaction opportunities for the third type of users to enhance their purchase conversion rates.

References

[1] Voronkova, V. H., Nikitenko, V. A., Teslenko, T. V., & Bilohur, V. E. (2020). Impact of the worldwide trends on the development of the digital economy. Amazonia investiga, 9(32), 81-90.

[2] Zekos, G. I., & Zekos, G. I. (2021). E-Globalization and Digital Economy. Economics and Law of Artificial Intelligence: Finance, Economic Impacts, Risk Management and Governance, 13-66.

[3] Yeo, S. F., Tan, C. L., Lim, K. B., Leong, J. Y., & Leong, Y. C. I. (2020). Effects of social media advertising on consumers' online purchase intentions. Global Business and Management Research, 12(1), 89-106.

[4] Anson, J., Boffa, M., & Helble, M. (2019). Consumer arbitrage in cross-border e-commerce. Review of international economics, 27(4), 1234-1251.

[5] Pileliené, L., Batyk, I. M., & Žukovskis, J. (2023). Cross-Border Shopping on the European Union Fast-Moving Consumer Goods Market: Determinants of Lithuanian Shoppers’ Behavior in Poland. Sustainability, 16(1), 102.

[6] Zhu, W., Mou, J., & Benyoucef, M. (2019). Exploring purchase intention in cross-border E-commerce: A three stage model. Journal of Retailing and Consumer Services, 51, 320-330.

[7] Lv, H. (2023). E-commerce consumer behavior analysis based on big data. Journal of Computational Methods in Science and Engineering, 23(2), 651-661.

[8] Wang, C., Liu, T., Zhu, Y., Wang, H., Wang, X., & Zhao, S. (2023). The influence of consumer perception on purchase intention: Evidence from cross-border E-commerce platforms. Heliyon, 9(11).

[9] Huang, S. L., & Chang, Y. C. (2019). Cross-border e-commerce: consumers' intention to shop on foreign websites. Internet Research, 29(6), 1256-1279.
[10] Han, B., Kim, M., & Lee, J. (2018). Exploring consumer attitudes and purchasing intentions of cross-border online shopping in Korea. Journal of Korea Trade, 22(2), 86-104.
[11] Cardona, M., Duch-Brown, N., & Martens, B. (2015). Consumer perceptions of cross-border e-commerce in the EU Digital Single Market (No. 2015/06). Institute for Prospective Technological Studies Digital Economy Working Paper.
[12] Xu, Y., He, D., & Fan, M. (2024). Antecedent research on cross-border E-commerce consumer purchase decision-making: The moderating role of platform-recommended advertisement characteristics. Heliyon, 10(18).
[13] Yang, Y., Yang, L., Chen, H., Yang, J., & Fan, C. (2020). Risk factors of consumer switching behaviour for cross-border e-commerce mobile platform. International Journal of Mobile Communications, 18(6), 641-664
[14] Huang, W. L., Hu, P., Tsai, S., & Chen, X. D. (2021). The business analysis on the home-bias of E-commerce consumer behavior. Electronic Commerce Research, 21, 855-879.
[15] Han, L., Ma, Y., Addo, P. C., Liao, M., & Fang, J. (2023). The role of platform quality on consumer purchase intention in the context of cross-border e-commerce: The evidence from Africa. Behavioral Sciences, 13(5), 385.
[16] Lu, Y., & Lv, S. (2025). Research on Cross-border E-commerce Consumer Behavior Based on Spatio-temporal Data Mining Calculation in the Perspective of Digital Economy Globalization. J. COMBIN. MATH. COMBIN. COMPUT, 127, 6711-6728.
[17] Yuwen, H., Guanxing, S., & Qiongwei, Y. (2022). Consumers' perceived trust evaluation of cross-border e-commerce platforms in the context of socialization. Procedia Computer Science, 199, 548-555.
[18] Chotisarn, N., & Phuthong, T. (2025). Logistics service quality and customer behavior in cross-border e-commerce: a thai consumer perspective. Cogent Business & Management, 12(1), 2486581.
[19] Zhang, H. (2021). STRATEGIES FOR ADDRESSING CONSUMER BEHAVIOR BARRIERS IN THE DEVELOPMENT OF CROSS-BORDER E-COMMERCE. Psychiatria Danubina, 33(suppl 8), 164-165.
[20] Yang, Y., & Lin, W. L. (2022). The Impact of Consumer Trust and Consumer Loyalty on Sustainable Development of Cross-border E-commerce. Special Education, 1(43).
[21] Muhammad Yaseen & Zara Karamat. (2025). Requirements Engineering Model (REM): An Assessment Model for Software Vendor Organizations. Journal of Software: Evolution and Process,37(4),e70020-e70020.
[22] Jianhua Wang & Nan An. (2024). Research on Evaluation Method of Green Suppliers Under Pythagorean Fuzzy Environment. Sustainability,16(20), 9124-9124.
[23] Lihua Liu. (2025). Application of K-means supported by clustered systems in big data association rule mining. Systems and Soft Computing,7,200211-200211.