Application of Improved Binary K-means Algorithm in Time and Cost Optimization for Regional Logistics Distribution Center Location

Dandan Wang
Guangxi Transport Vocational and Technical College, Nanning 530216, China
E-mail: wdd103603@163.com

Keywords: binary K-means algorithm, logistics distribution center, optimize site selection, time optimization, lowest cost

Received: September 26, 2024

The surge in express delivery volume has heightened the importance of addressing customer distribution needs. As a critical component of the logistics supply chain, the regional logistics distribution center requires strategic site selection to enhance service quality and reduce operational costs. This study proposes an optimized location model for regional logistics distribution centers based on an improved binary K-means clustering algorithm, focusing on minimizing distribution time and enterprise costs. In the initial stage, Z-score normalization was applied to preprocess the data and eliminate dimensional effects. The initial cluster centers were selected randomly from demand points to mitigate the risk of local optima. The model employs the time spent on the journey as the primary objective function while also incorporating enterprise investment cost and operational risk cost. With an initial setting of 9 cluster centers and a maximum of 100 iterations, the model demonstrated rapid convergence. Experimental results indicate that the total distribution time was reduced to 18,800 minutes, representing a 33.6% decrease compared to the conventional K-means model. The optimized solution identified 6 distribution centers, with the average distance from each center to demand points maintained below 1 km, effectively lowering enterprise costs and risks. The findings highlight the efficiency of the improved binary K-means clustering algorithm in optimizing time and cost, providing valuable insights for strategic site selection of regional logistics distribution centers.

Povzetek: Predlagan je izboljšan binarni K-means algoritem za optimizacijo lokacije regionalnih logističnih centrov, ki zmanjšuje stroške in čas dostave.

1 Introduction

The developed e-commerce model has led to the dramatic increase in the volume of domestic express delivery business. In the logistics system, the regional distribution center is a key node in the logistics network, connecting the various links in the supply chain and playing the role of the top and bottom [1]. Considering the importance of the distribution center in the logistics system, its location is particularly important. Reasonable site selection can significantly reduce the operating costs of enterprises, improve service efficiency, enhance customer satisfaction, so as to make enterprises develop stably [2]. The location selection of logistics distribution centers needs to consider many factors such as the storage capacity, distribution capacity, number of customer items, number of subordinate distribution points, distribution distance, distribution time, future business volume, etc. Traditional location selection methods usually implement decisions based on subjective experience, requiring decision-makers to have rich professional knowledge and experience. However, for different decision-makers, due to individual cognitive differences, Different decision-makers often obtain different decision outcomes. However, the construction of logistics distribution centers requires huge investment, and inappropriate decisions can lead to significant resource waste. Therefore, traditional subjective decision-making has significant differences and risks. And with computer developing continuously, researchers try to use data mining technology to help enterprises solve the problem of logistics distribution center location [3]. For logistics distribution site selection, the main thing is to gather the grass-roots distribution outlets into different clusters, so that the distance from each point in the cluster to the cluster center is the shortest [4]. The selected cluster center is the distribution center, which has the shortest distance and the lowest cost when distributing goods. The traditional K-means algorithm uses Euclidean distance to measure the similarity between points, with the shortest Euclidean distance as the optimization indicator. This approach is consistent with the demand that the shorter the actual delivery distance, the better. However, in actual delivery, the actual route between the distribution network and the distribution center is not a straight line. Therefore, obtaining the location of the distribution center through Euclidean distance is only the theoretical optimal position, not the optimal location for actual delivery; In addition, in actual delivery, in addition to the distance, delivery time is also an important consumer demand for customers. The same distance can lead to different delivery times in different road congestion situations. Therefore, when choosing a

logistics distribution center address, delivery time also needs to be considered. Based on this, this study constructs an improved K-mean regional distribution center optimization site selection model based on distribution time optimization in order to achieve the optimization of regional logistics center site selection with the objective of minimum distribution time. With the objective of minimum enterprise input cost and minimum enterprise business risk cost, the bifurcated K-mean clustering algorithm is used to solve the problem. The study aims to provide effective guidance for enterprises in selecting distribution centers.

This research is divided into four parts, the first part is the research results of domestic and foreign experts and scholars on logistics and distribution site selection and K-mean clustering algorithm in site selection optimization. The second part takes the minimum distribution time, the minimum enterprise input cost and the minimum enterprise business risk cost as the objectives, and adopts the improved K-mean regional distributing center optimization site selection model with optimizing distribution time and the dichotomous K-mean clustering algorithm for solving respectively. In the third part, the two models are experimented and analyzed, and in the fourth part, the article is summarized and deficiencies are pointed out.

2 Related works

The high-speed development of e-commerce has led to changes in consumer demand, and the optimization of logistics and distribution site selection is also deepening, and some experts and scholars have relevant research results in this regard. Liu et al. proposed a blockchain based method for selecting the location of agricultural product logistics distribution centers to overcome the problems of low delivery rate and high cost. This method combines blockchain technology to construct a site selection model for agricultural product logistics distribution centers based on input-output ratio. The mixed particle algorithm is used to solve the site selection model for agricultural product logistics distribution centers and obtain the optimal site selection scheme. The experimental results show that the on-time delivery rate of this method can reach 97.4% [5]. In order to solve the problem of e-commerce logistics distribution site selection, Shen constructed a time window logistics distribution site selection optimization model based on the dynamic uncertainty of the urban road network, while optimizing the path. The model will be solved using an improved genetic algorithm with minimum cost as the objective function. Example verification shows that this model can significantly reduce the logistics delivery time of e-commerce enterprises, thereby improving customer service satisfaction [6]. In order to achieve the goal of effective energy conservation and emission reduction in the cold chain logistics site selection process of fresh agricultural products, Wang optimized the product freshness and carbon emission target functions based on the original location of the distribution center. The constructed model is a dual objective function localization model that minimizes total cost and carbon emissions, which uses a two-stage heuristic function for solution. The proposed cold chain logistics site selection model for fresh agricultural products can effectively reduce logistics costs through the analysis of a certain enterprise site selection case [7]. Liu et al. to solve the optimal allocation of transshipment centers, processing plants, and distribution centers in the supply chain network under the conditions of uncertainty of transportation cost and customer demand. Considering the uncertainty of the supply chain, a two-stage fuzzy 0-1 mixed integer optimization model was developed. Considering the complexity of this model, an improved hybrid second-order particle swarm optimization algorithm is proposed to solve the resulting model. By comparing this model with the hybrid genetic algorithm, it is verified that the computational time and convergence speed of the model are optimal [8].

With the large increase in the amount of logistics and distribution data, K-mean clustering algorithm as a classic data mining technique is also applied to the logistics and distribution site selection optimization problem. Scholars at home and abroad have conducted extensive and in-depth research on the application of K-means clustering algorithm to logistics center location selection. Prabhu et al. proposed a supplier logistics location optimization method based on K-means clustering algorithm to reduce the transportation cost of logistics centers. The data was preprocessed using Z-score normalization in the optimization method, and the supplier logistics were grouped using K-means clustering algorithm. The example analysis results show that this method is effective in optimizing the location of logistics centers [9]. Yong believes that the location selection of distribution centers is related to the long-term development and stability of e-commerce. He has constructed a location optimization model to address the problem of low efficiency in traditional location selection algorithms. This model considers the accuracy of distribution locations and solves it using K-means clustering algorithm and improved particle swarm optimization algorithm. Through real data simulation experiments of logistics enterprises, it is known that the model greatly improves computational efficiency and positioning accuracy, and both logistics enterprises and customers are relatively satisfied with the delivery process [10]. Luo et al. to solve the unstable site selection effect of the traditional K-mean clustering algorithm, they proposed a K-mean algorithm based on the density of the neighborhood. The grid distribution characteristics of the samples are obtained through multidimensional grid division, while an iterative factor is introduced to merge the adjacent high-density grids to obtain a set of candidate initial clustering centers. Finally, the combination of grid density and distance is used to achieve distribution center location selection. The method has high accuracy and stability [11]. Liu et al. believe that the location problem of logistics distribution centers is a multi-attribute group decision-making problem that considers multiple product preference weights. They propose a K-clustering analysis method based on two-dimensional language similarity to improve the operation rules. The effectiveness and rationality of this method in

selecting distribution center locations have been verified through case analysis [12]. The above literatures are summarized in Table 1.

Table 1 Summary of relevant work

Researcher	Method	Experimental Results	Limitations
Liu et al. [5]	Blockchain + Hybrid Particle Algorithm	On-time delivery rate reached 97.4%	High computational complexity, limited applicability
Shen [6]	Genetic Algorithm, Dynamic Time Window	Enhanced effectiveness of location model	Neglects changes in customer demand in dynamic environments
Wang et al. [7]	Bi-objective Function + Heuristic Algorithm	Reduced logistics cost and carbon emissions	Complex model, long computation time
Liu et al. [8]	Improved Hybrid Second-order Particle Swarm Optimization	Optimal convergence speed and computation time	Applicability in complex environments needs validation
Prabhu et al. [9]	Z-score Normalization + K-means Clustering	Effectively reduced transportation costs	Less effective with large-scale data
Yong and Lu [10]	Improved Particle Swarm Optimization + K-means Clustering	Increased computational efficiency and location accuracy	Highly dependent on initial parameter selection
Luo et al. [11]	Neighborhood Density-based K-means Clustering	Improved stability and accuracy of location selection	Inefficient handling of dense area data
Liu and Li [12]	K-clustering Analysis Based on 2D Linguistic Similarity	Addressed multi-attribute group decision-making problems	Requires complex calculation rules, high computational cost

To sum up, in the optimization research of logistics site selection scheme, there is less research on time cost and lack of research on the site selection optimization of regional logistics and distribution centers. The application of K-mean clustering algorithm to the study of site selection optimization problem mainly combines the algorithm with other algorithms, and there is a lack of improvement of K-mean clustering algorithm separately. Therefore, this study constructs an improved K-mean regional distribution center optimization site selection model based on delivery time optimization using the distance spent time as the optimization objective. The input cost of the enterprise and the cost of the business risk of the enterprise are taken as the optimization objectives, and the bifurcated K-mean clustering is applied to solve the objective function. The paper aims to provide a reference scheme for the location selection of regional logistics centers.

3 Construction of regional logistics and distribution site selection objective optimization model

To optimize regional logistics distribution centers' location, this chapter is divided into two parts to construct an objective optimization model. Regional logistics distribution centers belong to the third level distribution centers, and logistics operations mainly include three categories: logistics exchange between the third level distribution center and the second level distribution center, logistics exchange between different regional logistics distribution centers, and logistics exchange between the third level distribution center and the fourth level distribution center. Design a location selection algorithm for regional logistics distribution centers based on the characteristics of logistics operations. Due to the high demand for delivery time in the business of regional logistics distribution centers, optimization of delivery time needs to be considered. At the same time, the investment cost of enterprises is the main factor that enterprise managers consider when formulating distribution site selection and construction plans. The investment cost considered in the study is the total cost of constructing the end point distribution network in the next year, including fixed costs and operating costs. The business risk cost considered in this article is due to the long distance from customers to the end of the distribution network for self pickup, resulting in low customer satisfaction and loss of some customers, thereby losing potential profits in the next year. The first part takes the travel time as the optimization objective, and constructs the improved K-mean regional distribution center optimization site selection model with optimizing distribution time. The second part takes the enterprise's input cost and the enterprise's business risk cost as the optimization objective, and uses the dichotomous K-mean clustering algorithm to solve the objective function.

3.1 Distribution time optimization model construction

Logistics distribution network has four levels. Firstly, it's the distribution center is to exchange the provincial logistics between these tasks [13]. The second level of logistics distribution center involves logistics

tasks between cities, the third level of distribution center consists of a number of regional distribution centers, the region is divided by the city logistics center business jurisdiction [14]. Regional distribution centers are the pillars of the distribution network, distribution system, with transit warehousing function [15]. Goods can first be transported to the regional distribution center through the trunk line, and then the goods will be distributed to various scattered destinations, such as retail outlets or consumers themselves. Regional logistics distribution centers are less affected by urban traffic rules, and the transportation used for distribution is mainly based on electric tricycles and minivans. The fourth level distribution center is some grass-roots outlets to realize the logistics distribution service [16]. Fig. 1 show the schematic diagram of the logistics distribution network.

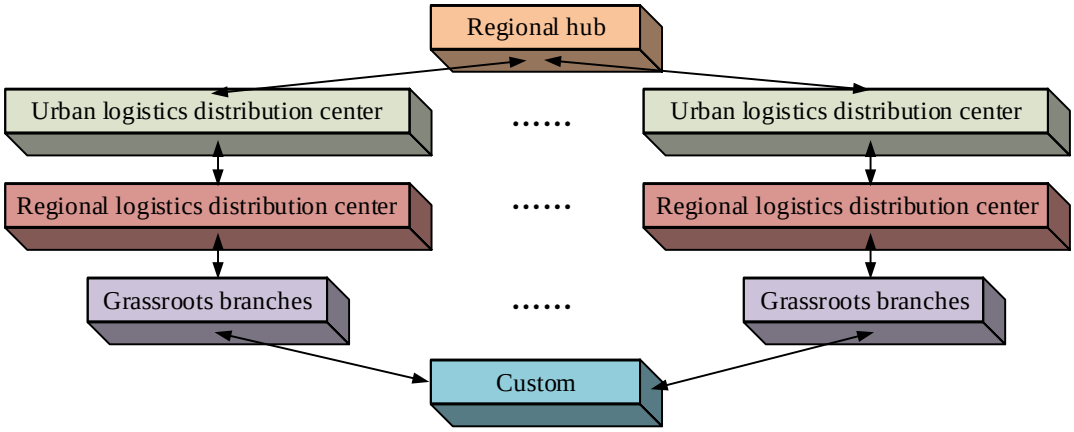


Figure 1: Schematic diagram of logistics distribution network

When the traditional K means clustering (K-means) is applied to the logistics distribution center location, the distribution distance is mainly calculated by L1 and L2 paradigms [17]. However, from the user's point of view, the delivery time is also a factor that should be taken into account. Therefore, to meet the customer's demand for distribution time, and reduce the time cost in logistics and distribution, an improved K-mean regional distribution center optimal location model based on distribution time optimization is constructed. Firstly, the Time Spent on the Single Journey (TSOTSJ) is defined as the objective function to measure the best candidate centers of clustered point clusters, and TSOTSJ can be expressed as the sum of intra-cluster delivery times of a cluster. And define a Time Spent on the Journey (TSOSJ) as an objective function to measure the quality of the final clustering result, which is the total distribution time of all clusters. Define a data set containing n data objects $D=\{x_1,x_2,\cdots x_n\}$, and the set of clusters generated by the traditional K-mean clustering algorithm is $Q=\{Q_1,Q_2,\cdots,Q_K\}$. Then define the data set of m data objects contained within the ith cluster as $Q_i=\{x_{i1},x_{i2},\cdots x_{im}\}$, and the formal definition of TSOTSJ is shown in Eq. (1).

$$TSOTSJ(i)=\sum_{j=1}^{m} f(Q_{ic}x_{ij}) \tag{1}$$

In Eq. (1), Q_{ic} denotes the center point of the cluster Q_i, x_{ij} denotes the points other than the center point in the cluster Q_i, and $f(Q_{ic}x_{ij})$ denotes the time spent from the point Q_{ic} to the point x_{ij}. According to Eq. (1), a set containing K optimal centroids can be defined as $C=\{Q_{1c},Q_{2c}\cdots Q_{Kc}\}$, which is generated by clustering the improved K-mean regional distribution center optimal site selection model based on distribution time optimization. Therefore, a formal definition of TSOSJ that measures the quality of the final clustering result can be obtained as shown in Eq. (2).

$$TSOSJ(C)=\sum_{i=1}^{K}\min(\sum_{j=1}^{m} f(Q_{ic}x_{ij})) \tag{2}$$

In Eq. (2), $\min(\sum_{j=1}^{m} f(Q_{ic}x_{ij}))$ represents the minimum intra cluster travel time of the i cluster. Eq. (1) measures the total delivery time within the coverage of a single distribution center and reflects the distribution efficiency in that area. By minimizing this time cost, delivery routes can be optimized and service timeliness improved. Eq. (2) is an extension of Eq. (1) and is used to calculate the total distribution time of all distribution centers as the core index to evaluate the overall logistics location and clustering effect. The optimization goal is to minimize the total distribution time, so as to improve the efficiency of the logistics system, reduce the time cost, meet the distribution needs in actual scenarios, and help enterprises improve operational efficiency and reduce costs.

Assuming that there are 10 points labeled A~J, the points first need to be clustered into 3 clusters and the optimal 3 centroids are obtained, i.e., the value of K is 3. Firstly, 3 clusters are obtained according to the traditional K-means, which are {A, F, I}, {B, D, E, G}, and {C, H, J}. Then the centroid of each cluster and the value of TSOTSJ of each cluster are calculated and checked whether the enumeration process of each cluster has been completed. Compare to get the smallest TSOTSJ value

that occurs during the enumeration process of each cluster, output the smallest TSOTSJ value of each cluster and get the final TSOSJ value. Finally, the centroid corresponding to the smallest TSOSJ value occurring in the self-enumeration process of each cluster is output, which is the selected distribution center site on the foundation of delivery time as the objective. The computational flowchart of the improved K-mean regional distribution center optimization site selection model on the foundation of delivery time optimization is shown in Fig. 2.

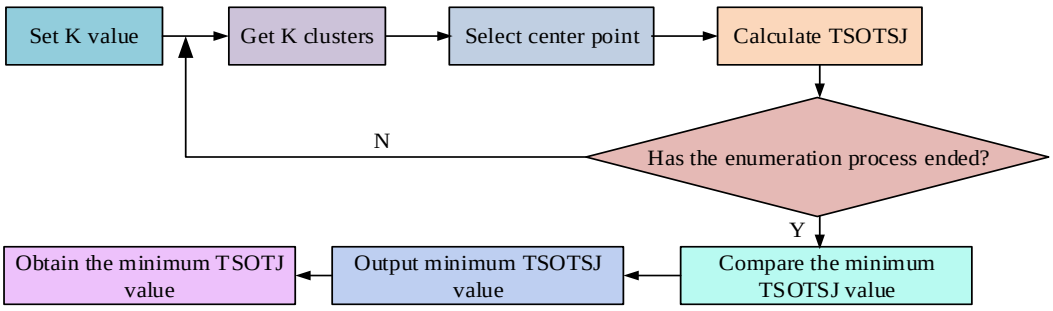


Figure 2: Flow chart of improved model calculation

3.2 Cost minimization optimization model construction

The input cost of the enterprise is also an important consideration for the location plan of the regional distribution center, which includes fixed costs and operating costs [18]. In addition to this, it is also necessary to consider the cost of business risk to avoid the loss of customers or distribution network operators due to the distance of customers or distribution outlets to the distribution center [19]. Fixed costs include one year's rent, procurement and installation costs of logistics facilities and equipment. Operating costs are mainly warehousing costs, equipment maintenance costs, human resource costs, utilities, etc. The main factor affecting this cost is the handling volume of goods, and if the handling turnover reaches a certain level it will create economies of scale [20]. Fig. 3 shows the cost schematic of enterprise.

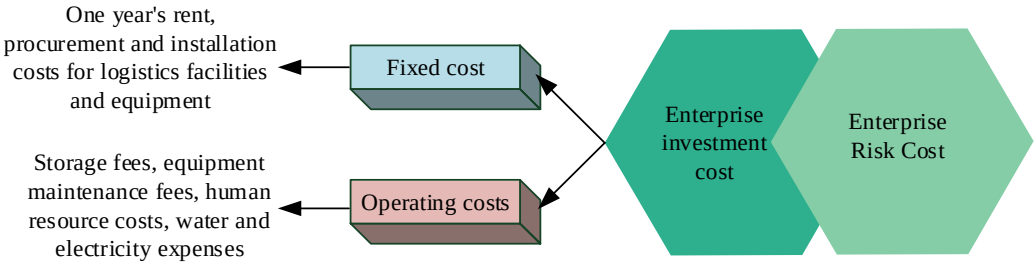


Figure 3: Cost diagram of the enterprise

The operation cost of a single piece of cargo will decrease with the increase of cargo handling volume until the upper limit of the handling capacity of the distribution center is reached. Based on this relationship, the distribution center operating cost function for different role volumes can be obtained as shown in Eq. (3).

$$v_i = a * \sqrt{g_i} \tag{3}$$

In Eq. (3), v_i represents the operation cost. g_i represents the operating volume of the i distributing center. a represents the coefficient. Business risk requires close attention to customer satisfaction in site selection, considering the relationship between customer satisfaction and pickup distance, we can get the relationship Eq. shown in Eq. (4).

$$D_j = \begin{cases} 1 & When\ d_j \le d_1 \\ \dfrac{(d_j - d_1)}{(d_2 - d_1)} & When\ d_1 < d_j < d_2 \\ 0 & When\ d_2 \le d_j \end{cases} \tag{4}$$

In Eq. (4), $[d_1, d_2]$ represents the range of acceptable pickup distance for customers, d_j represents the distance from the demand point j to the nearest distribution center, and D_j represents the satisfaction level of customers at the j demand point to pick up the goods at the distribution center. According to Eq. (4), the customer dissatisfaction can be obtained as $1 - D_j$, so the profit lost by customer churn is shown in Eq. (5).

$$Z = c_1 * c_2 * \sum (1 - D_j) * c_j \tag{5}$$

In Eq. (5), Z represents the cost of operational risk, c_1 represents the proportion of self-pickup customers, c_2 represents the average profit per customer per year, and c_j represents the number of customers at the j customer demand point. Considering the minimum input cost and the minimum operation risk cost, the objective function of both can be obtained as shown in Eq. (6).

$$\begin{cases} MinF = \sum (y_i + v_i) \\ MinZ = c_1 * c_2 * \sum (1 - D_j) * c_j \end{cases} \quad (6)$$

In Eq. (6), y_i represents the first i distributing center's fixed cost. Assuming that the maximum daily operation volume of this distributing center is g_1 and the express collection and delivery volume of each distribution center is not greater than g_1, the capacity constraints of the distribution center are shown in Eq. (7).

$$\sum_{j=0}^{N} x_{ij} \le g_1 \quad (7)$$

In Eq. (7), x_{ij} indicates whether the j customer demand point goes to the i distribution center to pick up the courier or send the shipment. Since a customer demand point can and can only be served by one distribution center, the service mode constraints of the distribution center can be obtained as shown in Eq. (8).

$$\sum_{i=1}^{I} x_{ij} = 1 \quad (8)$$

The fixed cost of a single distribution center is determined by the size, so the fixed cost constraint can be obtained as shown in Eq. (9).

$$y_i = \begin{cases} w_1 \quad g_2 < \sum_{j=0}^{N} x_{ij} \le g_1 \\ w_2 \quad g_3 < \sum_{j=0}^{N} x_{ij} \le g_2 \\ w_3 \quad g_4 < \sum_{j=0}^{N} x_{ij} \le g_3 \\ \cdots \\ w_k \quad \sum_{j=0}^{N} x_{ij} \le g_k \end{cases} \quad (9)$$

In Eq. (9), w_k represents the fixed cost of building a distribution center with the size of k, N refers to the total number of demands. The input cost for the construction of the regional distribution center should not exceed the enterprise's budget, so the construction cost constraint can be obtained as shown in Eq. (10).

$$F < A \quad (10)$$

In Eq. (10), F denotes the input cost of the enterprise, and A denotes the input cost budget of the enterprise. The decision variable x_{ij} in all the above equations can only take the value of 0 or 1, and all the variables should take the value greater than zero, and the values of the variables are shown in Eq. (11).

$$\begin{cases} i = 1, 2, 3, 4, \cdots \\ j = 1, 2, 3, 4, \cdots \\ k = 1, 2, 3, 4, \cdots \end{cases} \quad (11)$$

At the same time, the total capacity of all regional distribution centers also needs to satisfy the total customer demand. The customer demand constraint is shown in Eq. (12).

$$\sum_{k=1}^{K} g_k \ge \sum_{j=1}^{N} n_j \quad (12)$$

In the above Equation, n_j represents the express demand at the j demand point. Customers served by the regional distribution center, the overall satisfaction also needs to be satisfied to reach more than 80%, the satisfaction constraint is shown in Eq. (13). The reason for setting the overall satisfaction level at 80% is considered to achieve the purpose of site optimization, with reference to the "2023 Latest Principles for Logistics Distribution Center Site Selection".

$$\frac{\sum_{j=1}^{N} x_{ij} * D_j}{\sum_{j=1}^{N} x_{ij}} > 0.8 \quad (13)$$

The optimal solution between all the constraints mentioned above has many contradictions and cannot be optimized at the same time. The distribution center closest to the customer will reduce the risk of enterprise operation, but it will also cause the increase of enterprise input cost [21]. Therefore, how to coordinate the contradiction between enterprise input cost and operation risk needs to be based on the actual demand, set different weighting coefficients to get the optimal solution [22]. Based on this, the linear weighting method is used to construct the evaluation function [23]. Set the weight of enterprise input cost as 1 and the weight of ω. ω The larger it is, the greater the business risk of the enterprise, and the evaluation function is shown in Eq. (14).

$$MinQ = F + w * Z \quad (14)$$

In Eq. (14), F denotes the input cost of the enterprise, Z denotes the risk cost of the enterprise operation, and w denotes the coefficient. Due to the wide variety of regional distribution center site selection points, presenting significant spatial aggregation. K-means can divide the dispersed demand points into different categories according to the distance criterion to realize the spatial division [24, 25]. Fig. 4 shows the flowchart of traditional K-mean.

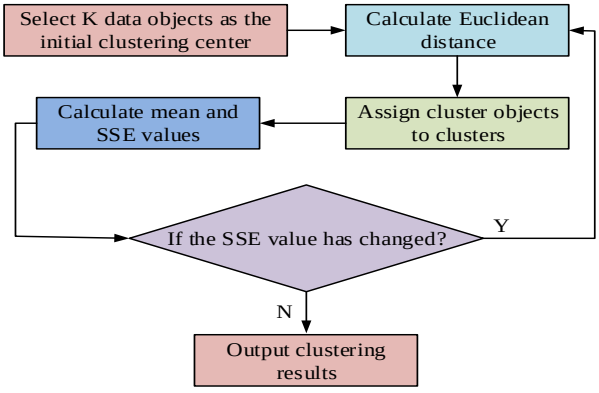


Fig. 4 Flow chart of traditional K-means

The current improvement of the K-means is mainly to improve the selection of the initial k-value, the removal of outliers and so on. The traditional K-means algorithm is easy to fall into local optimal solution when selecting the k value [26, 27]. Therefore, this article uses the bisection K-means to reduce the error that exists in the selection of the center of mass of the traditional algorithm and improve the operational efficiency. In solving the problem using the bifurcated K-means algorithm, each customer aggregation point is first regarded as a demand point, and the demand quantity of each demand point is known. Among all the demand points, two points are randomly selected as initial class centers. Assign all demand points to the nearest class centers to form different clusters. Find a new class center for each cluster, Eq. (15) show the location of the new class center.

$$\begin{cases} x = \dfrac{\sum x_i * c_i}{\sum c_i} \\ y = \dfrac{\sum y_i * c_i}{\sum c_i} \end{cases} \tag{15}$$

In the above Equation, (x, y) denotes the coordinates of the new class center within the cluster, (x_i, y_i) denotes the coordinates of the i demand point within the class, and c_i denotes the demand of the i demand point within the class. After many iterations, until the position of the cluster center within the cluster no longer changes or until the maximum number of iterations is reached. Fig. 5 shows the flowchart of the bisection K-means.

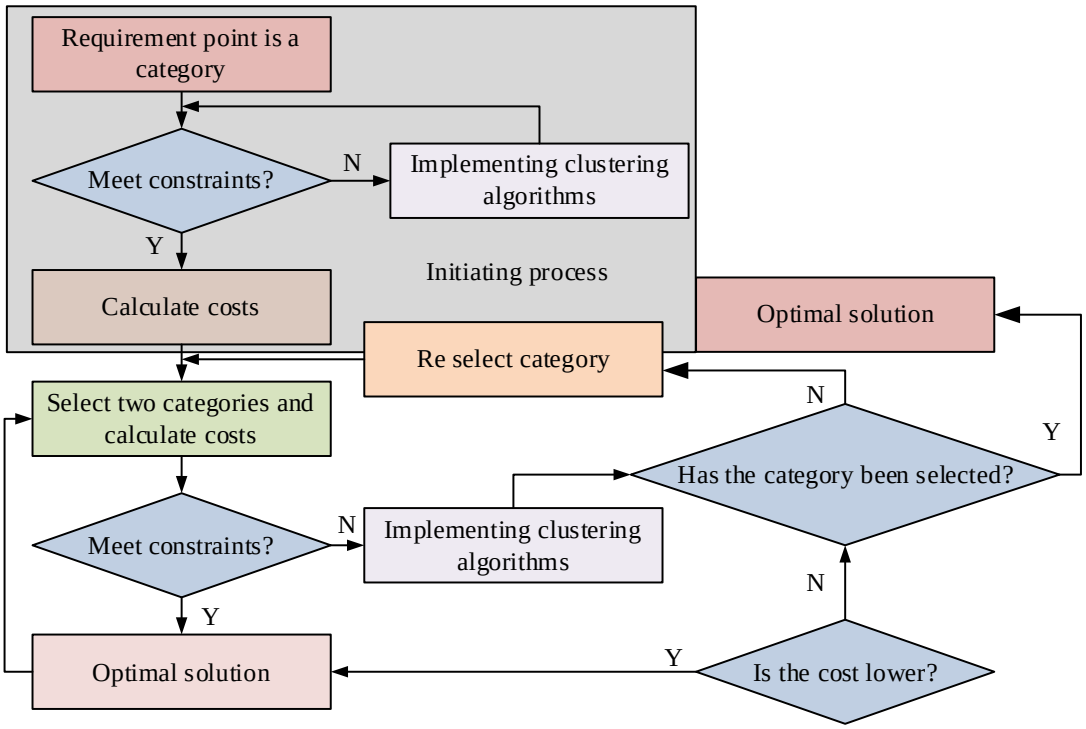


Fig. 5 Flow chart of the binary K-means

4 Model experimentation and analysis

To investigate the effectiveness of this optimal site selection model constructed for distribution time and cost in site selection, this chapter is divided into two parts to test the two models. The first part conducts experiments and analysis on the delivery time optimal site selection model. The second part conducts experiments and analysis on the least-cost optimal site selection model.

4.1 Delivery time optimization site selection model experiments and analysis

The experiments were performed on a high-performance computer with an Intel Core i9 processor, 32 GB of RAM, and an NVIDIA RTX 3080 graphics card to ensure efficiency for large-scale data processing. The data comes from 257 logistics outlets in a city, including the latitude and longitude of the outlets and the daily demand. In order to improve the clustering effect, Z-scores were standardized before the experiment, outliers were removed, and 9 initial clustering centers were randomly selected. The improved binary K-means algorithm sets the maximum number of iterations to 100, the convergence threshold to 10-4, and adopts Euclidean distance as the clustering criterion to better reflect the spatial relationship and delivery time cost in the logistics scene. Firstly, the sensitivity of the improved binary K-means clustering algorithm to key parameters, namely cluster number and weight coefficient, is studied and analyzed, and the stability and performance changes of the model under different parameter Settings are evaluated. Set the number of clusters to change from 5-15, and select the weight

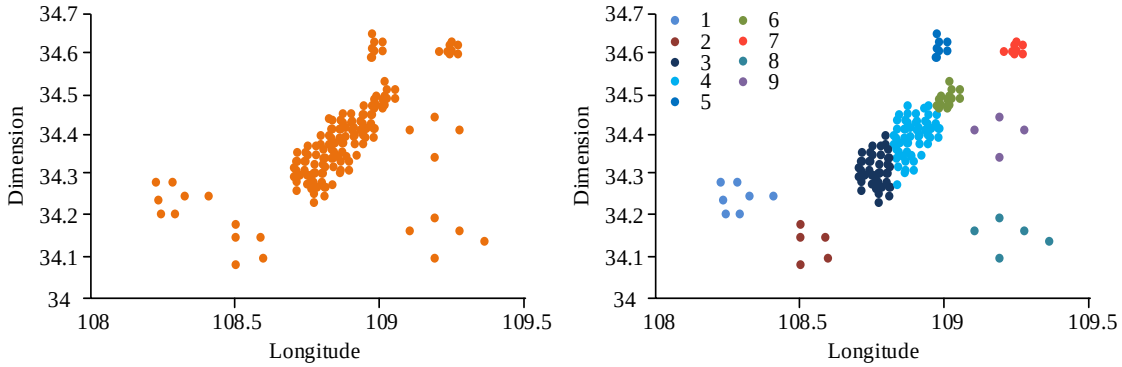
coefficient to change from 0.1-0.9. The results are shown in Table 2.

Table 2: Experimental results of sensitivity analysis of parameters

Cluster Number	Total Delivery Time (minutes)	Enterprise Cost (10,000 yuan)	Silhouette Score
5	25,400	3.20	0.58
7	21,200	2.90	0.63
9	18,800	2.45	0.72
12	19,000	2.50	0.70
15	20,500	2.64	0.65
Weight Coefficient	Time Cost (minutes)	Enterprise Cost (10,000 yuan)	Overall Optimization Effect
0.1	24,500	2.20	Bias Towards Enterprise Cost
0.3	22,000	2.30	Bias Towards Enterprise Cost
0.5	19,500	2.45	Best Balance
0.7	18,800	2.60	Bias Towards Time Cost
0.9	18,500	2.75	Bias Towards Time Cost

As can be seen from Table 2, with the cluster number increasing from 5 to 9, the total delivery time and enterprise cost will gradually decrease, and the Silhouette Score will also significantly improve, indicating improved clustering effect. The best results appear when the cluster number is 9, the total distribution time is 18,800 minutes, the enterprise cost is 2.45 million yuan, and the contour coefficient reaches 0.72, indicating the best tightness and separation of clustering results. When the cluster number continues to increase to 12 and 15, although the contour coefficient decreases, the changes of enterprise cost and total distribution time tend to be stable or even slightly increase, indicating that excessive cluster number will increase the calculation cost and the effect is not significantly improved. Therefore, clusters between 9 and 12 work best. The change of weight coefficient has obvious effect on the balance between time cost and enterprise cost. When the weight coefficient value is small (0.1 and 0.3), the model is more inclined to optimize the enterprise cost, but the time cost is higher. When the weight coefficient value is large (0.7 and 0.9), the model tends to minimize the time cost, but the enterprise cost increases. When the optimal weight coefficient is 0.5, the balance between time cost and enterprise cost is reached, the total delivery time is 19,500 minutes, and the enterprise cost is 2.45 million yuan. At this time, the model shows the best comprehensive optimization effect, which is suitable for most logistics distribution scenarios.

Using distance as the clustering criterion, the distribution map of grassroots outlets in the city as well as the clustering effect are shown in Fig. 6.



(a) Grassroots Network Distribution Map (b) Cluster rendering

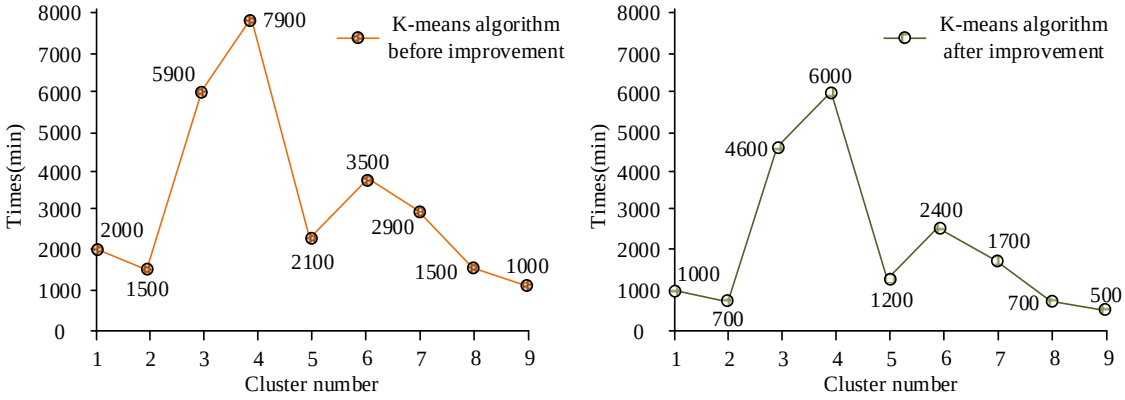
Figure 6: Grassroots network distribution map and clustering effect

Fig. 6(a) is the distribution of grassroots outlets in the city and Fig. 6(b) is the clustering effect. The regional outlets show a centralized distribution within the longitude of (108.8, 109.1) and the dimension of (34.25, 34.51), with sporadic distribution in other locations. Using distance as the clustering criterion, the clustered grassroots outlets were divided into 9 clusters. The number of outlets in the clusters numbered 1-9 are 7, 5, 80, 120, 8, 18, 10, 5 and 4, respectively. The specific latitude and longitude coordinates of the addresses of the regional logistics and distribution centers obtained by this algorithm are shown in Table 2.

Table 2: Specific longitude and latitude coordinates of the distribution center

Number	Latitude and longitude coordinates
1	(108.25, 34.18)
2	(108.50, 34.15)
3	(108.75, 34.30)
4	(108.80, 34.35)
5	(109.00, 34.62)
6	(109.00, 34.50)
7	(109.20, 34.60)
8	(109.30, 34.15)
9	(109.20, 34.19)

The specific coordinates of the addresses of the regional logistics and distribution centers for the nine clusters are given in Table 2. The above coordinate points are obtained with the objective of minimum path, i.e., minimum travel time. The TSOTJ values of the total travel time spent for the 9 clusters calculated by the K-means before and after the improvement are compared in Fig. 7.



(a) Algorithm before improvement (b) Algorithm after improvement

Figure 7: Comparison of TSOTJ values before and after improvement

Fig. 7(a) shows the TSOTJ values calculated by the pre-improved K-means algorithm and Fig. 7(b) shows the TSOTJ values calculated by the improved K-means. The sum of TSOTJ values calculated by the pre-improved K-means algorithm is 28,300 minutes, and the sum of TSOTJ values calculated by the improved K-means algorithm is 18,800 minutes, which is a 33.6% reduction in time. Therefore, it is proved that the improved K-means regional distributing center optimal siting model on the foundation of distribution time optimization can make the time cost reduced.

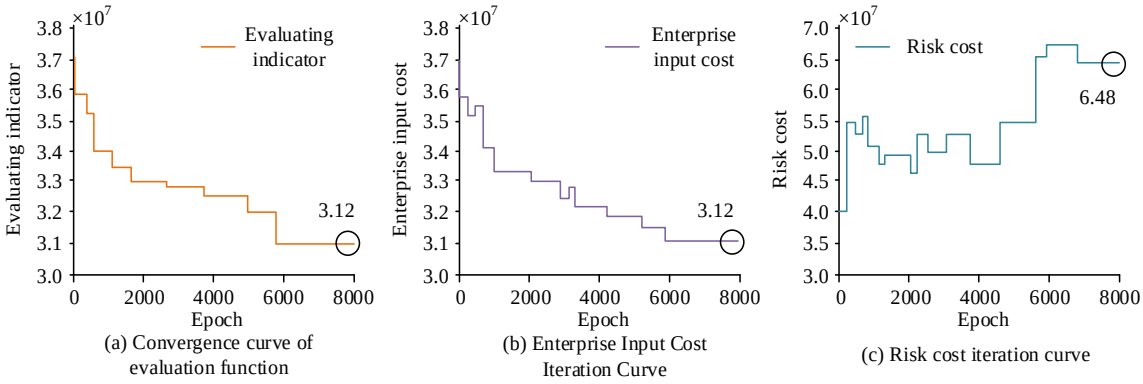
4.2 Experiments and analysis of the least-cost optimal site selection model

Take a logistics enterprise in a city as an example, it is proposed to establish four sizes of regional distribution centers in the city. After the research, it is known that the area of different sizes of regional distribution centers, the number of pieces of daily operations, and fixed costs are shown in Table 3.

Table 3: Cost table for regional distribution centers

Distribution center size	Area $(m)^2$	Daily workload (pieces)	Fixed cost (10,000 *yuan*)
Small scale	80	0-5000	35
Medium scale	120	5000-8000	55
Mass	160	8000-10000	72
In super-large scale	200	10000-12000	84

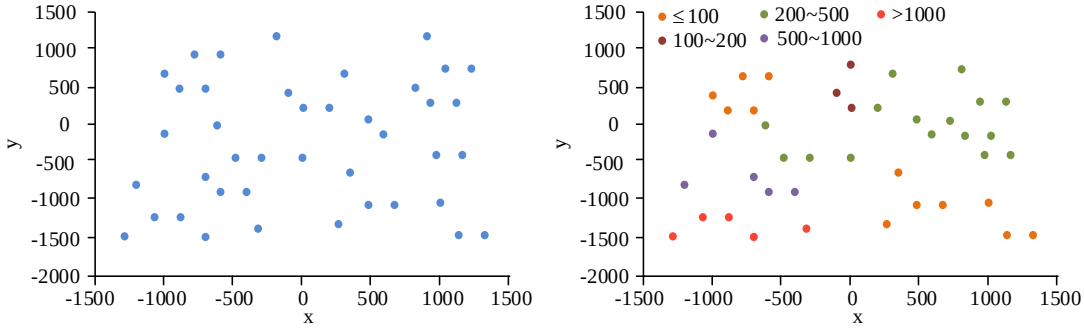
When selecting the location of the regional distribution center, it is necessary to consider the long-term operation of the enterprise and reduce the operational risk. When converting the multi-objective function to a single objective for solving, the weight is set to 1. The initial solution is found by using the K-means and its dichotomous form, and the convergence curves of the iteration of the evaluation function, the iteration curves of the enterprise's input cost and risk cost are shown in Fig. 8.



(a) Convergence curve of evaluation function (b) Enterprise Input Cost Iteration Curve (c) Risk cost iteration curve

Figure 8: Evaluation function and cost iteration curve

From Fig. 8(a), the evaluation function gradually decreases during the iteration process of optimization search until convergence, which proves the effectiveness of this algorithm. From Fig. 8(b), the input cost of the enterprise shows an overall decreasing trend in the iterative process, but there is a local increase, which is due to the fact that the evaluation function is determined by the enterprise's input cost and the cost of risk together. When the evaluation function decreases, the input cost is increased in order to reduce the risk of enterprise. From Fig. 8(c), there are up and down fluctuations in business risk, which is due to the magnitude of business input costs compared to the magnitude of business risk costs. Based on the map and research the coordinates and demand of different base outlets in a region of the city are derived as shown in Fig. 9.



(a) Coordinates of grassroots network points (b) Demand for grassroots branches

Figure 9: Coordinates and demand of grassroots network points

Fig. 9(a) shows the distribution of coordinates of different base outlets in a region of the city and Fig. 9(b) shows the demand of different base outlets in a region of the city. It can be seen that there are a total of 12 outlets in the region with express demand ≤ 100 and 3 outlets with demand between 100 and 200. There are 16 outlets with demand between 200 and 500, 5 outlets with demand between 500 and 1000, and 5 outlets with demand > 1000. The solution is performed according to the bisection K-means algorithm, and the coordinates of the distribution center are found as shown in Fig. 10.

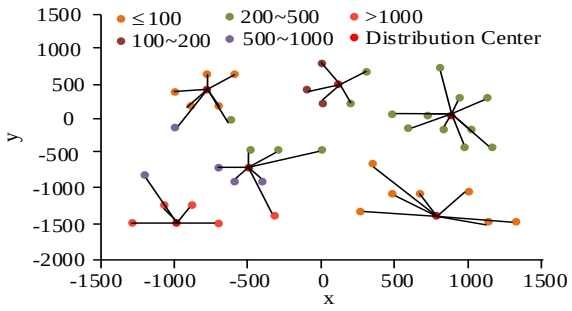


Figure 10: Distribution center distribution map

In Fig. 10, a total of six distribution centers are planned, with a total of three small distribution centers and three medium-sized distribution centers. Each grassroots outlet is distributed by the nearest distribution center, and the distance of customer pickup or distribution is within 1 km, which meets the needs of customers and the person in charge of the outlet.

In order to verify the effectiveness and rationality of the algorithm for solving the optimization problem of regional logistics distribution site selection, the latest regional logistics distribution site selection optimization research algorithm will be selected for site selection effect verification. The specific results are shown in Table 4. The latest research algorithms for optimizing regional logistics distribution location include K-means clustering algorithm, Emperor Butterfly optimization algorithm, improved multi-objective genetic algorithm, and ant colony immune algorithm. Compared with other regional logistics distribution location optimization algorithms, under four different area sizes, there are more daily tasks, lower fixed costs, and shorter delivery times. The range of daily homework items is 0-12000, with a fixed cost of 350000-840000 yuan and a delivery time range of 14-48

minutes. The K-means clustering algorithm and the Emperor Butterfly optimization algorithm also have significant advantages in three indicators: daily homework quantity, fixed cost, and delivery time. This indicates that the regional logistics distribution location optimization algorithm proposed by the research institute has more efficient time and cost optimization effects, and has more obvious advantages in practical applications such as low-carbon logistics distribution.

Table 4: Comparison of application effects of algorithms for optimizing regional logistics distribution site selection

Distribution center size	Area(m^2)	Daily workload(pieces)	Fixed cost (10000 yuan)	Delivery time/min
Research method	80	0-5000	35	14
	120	5000-8000	55	26
	160	8000-10000	72	38
	200	10000-12000	84	48
K-means clustering algorithm and emperor butterfly optimization algorithm	80	0-4000	41	17
	120	5000-6000	58	30
	160	6000-9000	75	42
	200	9000-10000	90	53
Improved multi-objective genetic algorithm	80	0-4000	43	19
	120	5000-7000	58	30
	160	7000-10000	74	41
	200	10000-11000	92	52
Ant Colony Immune Algorithm	80	0-5000	46	21
	120	5000-7000	61	34
	160	7000-9000	79	46
	200	9000-10000	95	54

5 Discussion

The improved dichotomy K-means clustering algorithm proposed in this study shows significant advantages in regional logistics distribution location optimization, and has higher efficiency and practical application value compared with the existing research. Traditional K-means clustering algorithm often faces problems such as high randomness in initial cluster center selection, easy to fall into local optimization, low efficiency of intensive data processing, especially in large-scale logistics distribution site selection, and high computing cost. While the neighborhood density optimization algorithm proposed by Luo et al. [11] has improved stability, its efficiency is limited when dealing with dense regions. The particle swarm optimization combined with K-means method adopted by Yong et al. [10] also has the problem of large dependence on initial parameter selection. In contrast, the Z-score standardized data preprocessing and the random selection of initial clustering centers are adopted in this study, which effectively avoids the local optimal problem and significantly improves the convergence speed of the algorithm. In addition, traditional researches, such as the genetic algorithm path optimization model adopted by Shen [6], mostly focus on the minimization of path distance and fail to fully consider the cost factors of enterprises. However, this study comprehensively considers the time cost, enterprise investment and operational risk cost, and builds a more comprehensive optimization framework.

The experimental results show that the proposed improved algorithm achieves rapid convergence after only 100 iterations, and the total delivery time is reduced by 33.6% from 28,300 minutes in the traditional K-means model to 18800 minutes. After optimization, the number of distribution centers is 6, and the average distance from each distribution center to the demand point is controlled within 1 km, which not only significantly reduces the time cost, but also effectively reduces the investment and risk of enterprises. Compared with high computational complexity schemes such as blockchain combined with hybrid particle algorithm, the improved binary K-means clustering algorithm is more outstanding in adaptability, computational efficiency and scalability, and can be flexibly applied to e-commerce logistics scenarios of different regions and data scales to meet the rapidly developing market demand. In summary, the method proposed in this study provides an efficient, economical and practical solution for logistics distribution site selection, which helps enterprises to improve service quality while reducing operating costs, and provides strong support for the development of the logistics industry.

6 Conclusion

In order to solve the problem of regional logistics distribution center location optimization, this study

constructs an improved K-mean regional logistics distribution location optimization model. Aiming at minimum enterprise input cost and minimum enterprise operating risk cost, binary K-means clustering algorithm is used to solve the problem. The experimental results show that the TSOTJ value calculated by the improved K-means algorithm is compared with the TSOTJ value calculated by the improved K-means algorithm. The total TSOTJ value calculated by the improved K-means algorithm is 28,300 minutes, and the total TSOTJ value calculated by the improved K-means algorithm is 18800 minutes, which reduces the time by 33.6%. The regional logistics distribution location optimization model constructed by the research can effectively reduce the time cost. In the function iteration process of the binary K-means algorithm, the loss value of the function gradually decreases and eventually tends to be stable, with a stable value of 3.12, which shows that the algorithm can find the initial solution. In the process of iteration, the overall input cost of the enterprise shows a downward trend, but there is a local increase, and the enterprise operation risk fluctuates up and down. According to the optimal solution of binary K-means, there are 6 distribution centers in total, and the distance between each distribution center and the demand point is less than 1 km.

In summary, the improved dichotomous K-means clustering algorithm proposed in the study has a good application effect in regional logistics distribution location. In the urban scene, the algorithm can deal with the high-density network distribution, optimize the distribution path, and reduce the impact of traffic congestion. In the rural scene, the algorithm dynamically adjusts the clustering center to effectively cover scattered nodes and improve efficiency. At the same time, the algorithm shows high computational efficiency when dealing with large-scale distribution networks, and can be flexibly adjusted to meet local needs in small-scale distribution networks. However, future research should further explore the comparative analysis of different clustering and optimization algorithms, and optimize with deep learning models to develop dynamic location models with real-time response capabilities. In addition, multi-objective optimization and data testing in different regions should also be considered to improve the robustness and universality of the model, and provide innovative support for logistics location decisions in different market environments.

Availability of data and material

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Competing interests

No conflict of interest exits in the submission of this manuscript.

References

[1] Z. Zhang. An optimization model for logistics distribution network of cross-border e-commerce based on personalized recommendation algorithm, Security and Communication Networks, 14(4): 1-11, 2021. https://doi.org/10.1155/2021/551048

[2] H. Zhao, A. Sharma. Logistics distribution route optimization based on improved particle swarm optimization, Informatica, 47(2): 243-252, 2023. https://doi.org/10.31449/inf.v47i2.4011

[3] T. Agajie, A. Salau, H. E. Abel. Power loss mitigation and voltage profile improvement with distributed generation using grid-based multi-objective harmony search algorithm, Journal of Electrical and Electronics Engineering, 13(2): 5-10, 2020.

[4] P. Chen, X. Zheng, F. Gu. Path distance-based map matching for Wi-Fi fingerprinting positioning, Future Generation Computer Systems, 107(6): 82-94, 2020. https://doi.org/10.1016/j.future.2020.01.053

[5] X. Liu, M. Luo. Site-Selection method of agricultural products logistics distribution center based on Blockchain. International Journal of Information and Communication Technology, 22(1): 15-31, 2023. https://doi.org/10.1504/IJICT.2023.127673

[6] Y. Shen. Optimization of urban logistics distribution path under dynamic traffic network, International Core Journal of Engineering, 6(1): 243-248, 2020. https://doi.org/%2010.6919/ICJE.202001_6,%20no.%201).0035

[7] H. Wang, H. Ran, X. Dang. Location optimization of fresh agricultural products cold chain distribution center under carbon emission constraints, Sustainability, 14(11): 1-24, 2022. https://doi.org/10.3390/su14116726

[8] Z. Liu, S. Qu, Z. Wu, Y. Ji. Two-Stage fuzzy mixed integer optimization model for three-level location allocation problems under uncertain environment, Journal of Intelligent and Fuzzy Systems, 39(5): 1-16, 2020. https://doi.org/%2010.3233/JIFS-191453

[9] R. M. Prabhu, G. Hema, S. Chepure, M. N. Guptha. Logistics optimization in supply chain management using clustering algorithms, Scalable Computing, 21(1): 107-114, 2020. https://doi.org/10.12694/scpe.v21i1.1628

[10] C. Yong, Y. Lu. An improved particle swarm optimization algorithm in selection of e-commerce distribution center, Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 39(6): 8783-8793, 2020. https://doi.org/10.3233/JIFS-189275

[11] M. Luo, Y. Yuan, K. Wang. An initial clustering center optimization method based on neighbourhood density for K-Means, Journal of Physics: Conference Series, 1748(3): 16-23, 2021. https://doi.org/10.1088/1742-6596/1748/3/032016

[12] P. Liu, Y. Li. Multiattribute decision method for comprehensive logistics distribution center location selection based on 2-Dimensional linguistic

information, Information Sciences, 538(10): 209-244, 2020. https://doi.org/10.1016/j.ins.2020.05.131

[13] Z. K. Hou. The optimization of automated goods dynamic allocation and warehousing model, Computer Optics, 44(5): 843-847, 2020. https://doi.org/10.18287/2412-6179-CO-682

[14] A. Strömer, N. Klein, C. Staerk, H. Klinkhammer, A. Mayr. Boosting multivariate structured additive distributional regression models, Statistics in Medicine, 42(11): 1779-1801, 2023. https://doi.org/10.1002/sim.9699

[15] H. Chen, H. Liu, B. Wu, H. Chen. An intelligent algorithm based on evolutionary strategy and clustering algorithm for lamb wave defect location, Structural Health Monitoring, 20(4): 2088-2109, 2021. https://doi.org/10.1177/1475921720959590

[16] E. Sanci, M. S. Daskin. An integer l-shaped algorithm for the integrated location and network restoration problem in disaster relief, Transportation Research Part B Methodological, 145(3): 152-184, 2021. https://doi.org/10.1016/j.trb.2021.01.005

[17] I. Fedorchenko, A. Oliinyk, J. A. J. Alsayaydeh. Modified genetic algorithm to determine the location of the distribution power supply networks in the city, Journal of Engineering and Applied Sciences, 15(23): 2850-2867, 2020.

[18] C. Ma, B. Li, J. He. The improved fault location method for flexible direct current grid based on clustering and iterating algorithm, IET Renewable Power Generation, 15(15): 3577-3587, 2021. https://doi.org/10.1049/rpg2.12246

[19] S. Wang, Y. Sun, Z. Bao. On the efficiency of k-means clustering: evaluation, optimization, and algorithm selection, Proceedings of the VLDB Endowment, 14(2): 163-175, 2020. https://doi.org/10.14778/3425879.3425887

[20] X. Zhu, S. Luo. The influence of computer network technology on national income distribution under the background of social economy, Computer Communications, 177(9): 166-175, 2021. https://doi.org/10.1016/j.comcom.2021.06.025

[21] X. Xiahou, Y. Harada. B2C E-Commerce customer churn prediction based on K-Means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 17(2): 458-475. 2022. https://doi.org/10.3390/jtaer17020024

[22] E. B. Tirkolaee, A. E. Torkayesh. A cluster-based stratified hybrid decision support model under uncertainty: sustainable healthcare landfill location selection, Applied Intelligence, 52(12): 13614-13633, 2022. https://doi.org/10.1007/s10489-022-03335-4

[23] P. Govender, V. Sivakumar. Application of K-Means and hierarchical clustering techniques for analysis of air pollution: a review (1980-2019), Atmospheric Pollution Research, 11(1): 40-56, 2020. https://doi.org/10.1016/j.apr.2019.09.009

[24] J. Zan. Research on robot path perception and optimization technology based on whale optimization algorithm, Journal of Computational and Cognitive Engineering, 1(4): 201-208, 2022. https://doi.org/10.47852/bonviewJCCE5978202055

[25] P. A. Ejegwa, J. M. Agbetayo. Similarity-Distance decision-making technique and its applications via intuitionistic fuzzy pairs, Journal of Computational and Cognitive Engineering, 2(1): 68-74, 2023. https://doi.org/10.47852/bonviewJCCE5125225

[26] D. Shi, G. Dong, E. Chen, M. Dai, N. Xiao, Y. Zhang, W. Chu. Optimization of storage paths for finished cigarette logistics distribution based on improved GA-A, Informatica, 48(18): 140-154, 2024. https://doi.org/10.31449/inf.v48i18.6436

[27] Y. Yang. The impact of GA optimization model under the constraint of maximum inventory on the logistics cost control of automotive parts production in the factory, Informatica, 48(11): 1-14, 2024. https://doi.org/10.31449/inf.v48i11.5959