IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 October 2023, accepted 4 January 2024, date of publication 12 January 2024, date of current version 23 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3353261

RESEARCH ARTICLE

Research on Fresh Pre-Positioning Warehouse Layout Based on Spatial Data Mining

WEI XU[1], CHAO WANG[1], LEI XING[1], AND NAN LI[2]

[1]College of Transportation, Shandong University of Science and Technology, Qingdao 266590, China
[2]Qingdao Branch of Shandong Road and Bridge, Qingdao 266100, China

Corresponding author: Lei Xing (xinglei0915@163.com)

This work was supported in part by the Joint Funding of the Shandong Social Science Planning Research Project 18CCXJ25, in part by the Qingdao Social Science Planning Research Project QDSKL1801134, in part by the Qingdao Shuangbai Research Project 2019-B-10, and in part by the Shandong Provincial Natural Science Foundation Youth Project ZR2023QG039.

ABSTRACT With the development of e-commerce logistics in China, online shopping has become increasingly popular among consumers, and the fresh e-commerce industry has shifted to a pre-positioning warehouse distribution model closer to consumers. This study starts from consumer demands and considers the spatial distribution characteristics of consumers. Through cluster analysis, demand points with high consumer density are identified, and potential locations are determined through two rounds of clustering. Drawing insights from the warehousing and distribution models of Freshippo and MissFresh in Wuhan and Nanjing, four classification algorithms are compared to determine the optimal pre-positioning warehouse distribution model for the enterprise in this study. A multi-objective optimization model is established, aiming to minimize enterprise costs and maximize customer satisfaction, and the optimal site selection plan is obtained. By determining the number and locations of pre-positioning warehouses through the optimal solution, the enterprise can achieve maximum investment returns in the spatial layout of these warehouses.

INDEX TERMS Pre-positioning warehouse layout, cluster analysis, classification algorithms, NSGA-II algorithm.

I. INTRODUCTION

With the rapid development of the national economy and e-commerce, online shopping has become an indispensable part of people's lives. The variety of products purchased online has been continuously increasing, and the market for fresh and other cold chain products in the e-commerce sector is rapidly growing. Consumers are placing greater emphasis on their shopping experience and the high-quality services provided by logistics companies. In order to meet consumers' demand for both speed and quality, fresh e-commerce has shifted towards a pre-warehouse distribution model located closer to consumers. The pre-warehouse model offers advantages such as reasonable inventory levels and short-distance delivery, which improve delivery efficiency, ensure freshness, and reduce losses.

Currently, a similar model to the pre-warehouse in foreign countries is the "ministore" model, which is primarily operated and managed by professional logistics companies. In China, the products distributed through pre-warehouses are mainly fresh produce. Companies such as Freshippo, JD Fresh, Dingdong Maicai, MissFresh, and Walmart's Sam's Club employ this model. Among them, MissFresh and Dingdong Maicai adopt the dark warehouse distribution model (pre-warehouses only provide storage functions without offline stores), while Sam's Club, Freshippo, and JD Fresh adopt a warehouse-store integrated model (pre-warehouses serve as both product storage and offline storefronts). Considering the functional value that pre-warehouses aim to achieve, customer demand is a crucial factor influencing the selection of pre-warehouse locations. By establishing the number and locations of pre-warehouses based on customer demand, both customer satisfaction and cost reduction for fresh food companies can be achieved. Therefore, this study clusters

The associate editor coordinating the review of this manuscript and approving it for publication was Kashif Munir.

customer demand points, establishes suitable warehouse types for clustering centers, and ultimately constructs a multi-objective optimization model that minimizes enterprise cost and maximizes customer satisfaction to provide reference for pre-warehouse site selection.

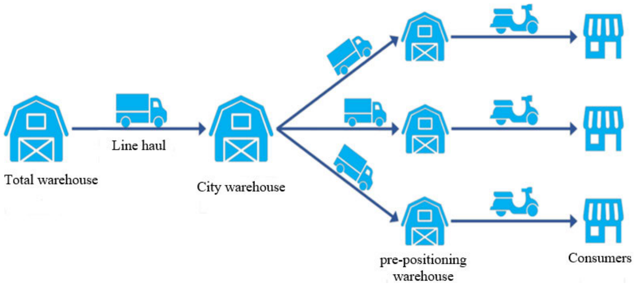


FIGURE 1. **Front-loading warehouse business model.**

II. LITERATURE REVIEW ON DATA MINING ALGORITHMS AND SITE SELECTION MODELS

A. DATA MINING ALGORITHMS

In the era of data-driven efficiency improvement, the application of spatial data mining methods in the research on layout and classification of fresh product pre-warehouses becomes more meaningful. Some scholars have applied clustering algorithms to location selection problems. Xiaoye et al. [1] introduced the K-means clustering method with non-linear coefficient of roads to construct a clustering model for pre-warehouse site selection. Liu et al. [2], combining GIS spatial analysis and clustering algorithms, identified candidate locations for solid urban waste landfills. Gocer et al. [3] improved the k-means clustering algorithm and combined it with multi-criteria decision-making to propose a new hub location selection method. In the research of improving clustering algorithms and utilizing clustering to enhance other algorithms, Haili [4] proposed an improved K-means algorithm based on the differential algorithm, using the standard deviation algorithm to determine the initial cluster centers in the K-means algorithm. Shuyang et al. [5] improved the initialization of the pheromone matrix in the ant colony algorithm based on the clustering partitioning by DBSCAN, thus enhancing the ant colony algorithm. Neeba et al. [6] proposed a hybrid intelligent algorithm combining PSO and clustering algorithms. Moreover, clustering algorithms are also applied to vehicle routing problems and surveillance and tracking of individuals [7], [8], [9], [10]. Currently, there is limited research on the classification of fresh product pre-warehouses. In other research fields, Weigang et al. [11] designed a weighted naive Bayes classification model to improve the accuracy and speed of ship trajectory classification. Yong et al. [12] selected ten years of garbage data from Shanghai and used a naive Bayes classifier to establish standards for the classification of household garbage. Yingzi et al. [13] addressed the problem of overfitting to large clusters in classifier-based news distribution with uneven distribution, proposing a weighted complement set-based naive Bayes algorithm model. Chen et al. [14] proposed an effective classification method using a weighted XGBoost model for complex radar signals. Wu et al. [15] established an XGBoost multi-class fault identification model for real-time fault diagnosis of wind turbines. Liu et al. [16] addressed the issues of high computational cost and overfitting in traditional forecasting methods, proposing a short-term power load forecasting method based on clustering and XGBoost algorithm. Song et al. [17] constructed an ecommerce consumer purchasing prediction model based on the XGBoost algorithm.

Most of the above studies are conducted under a single clustering or classification algorithm. Although these studies have achieved certain results in different problems, they have not compared the effects of different clustering algorithms on processing feature datasets for pre-warehouse layout, nor compared the capabilities of machine learning and ensemble learning in processing various heterogeneous data. Therefore, this paper adopts a two-stage clustering approach to handle noise points, and conducts comparative analysis on the same dataset with different clustering algorithms, clustering the customer demand points that are scattered in the demand area. At the same time, this study explores multiple heterogeneous datasets, consisting of text and continuous data, to compare the abilities of machine learning and ensemble learning in handling heterogeneity, in order to determine the warehouse types that are suitable for clustering center points.

B. INTRODUCTION TO SITE SELECTION MODELS

The fundamental purpose of fresh product warehouse facility layout is to reduce costs. Scholars have constructed site selection models from the perspective of cold chain logistics costs to improve the enterprise cold chain logistics system. Zhoufang et al. [18] used the principle of minimum distribution cost and applied clustering ideas and ant colony algorithms to solve the location selection problem of distribution centers. Yong [19] built an LTC site selection model for fresh ecommerce O2O community stores consisting of facility costs, transportation and distribution costs, management and operation costs, and customer satisfaction loss costs. Zhenbo et al. [20] combined the existing regional distribution center warehouse layout of H ecommerce, using the total logistics cost of transportation cost, warehousing operation cost, and return cost as the objective function, and adopted mixed integer programming to establish a mathematical model. Shiyu et al. [21] improved the simulated annealing algorithm to determine suitable solutions that minimize operating costs such as distribution and warehousing costs, effectively improving the efficiency of logistics distribution. Mingwei [22] established a bi-level programming site selection model with the upper level aiming to minimize enterprise costs and the lower level aiming to maximize consumer utilities, and used heuristic algorithms to solve the model. Weng [23] considered the cost of cargo damage and

TABLE 1. Symbolic description of the model.

Parameter	Description	Unit
c_{ij}	Transportation cost per unit of cargo volume per unit of distance from the *i*th large warehouse to the *j*th forward warehouse	yuan
c_{jk}	Transportation cost per unit of cargo volume per unit of distance from the *j*th forward warehouse to the *k*th customer	yuan
x_{ij}	Transportation volume of goods from the *i*th warehouse to the *j*th warehouse	kg
y_{jk}	Transportation volume of goods from the *j*th warehouse to the *k*th customer	kg
d_{ij}	Transportation distance from the *i*th warehouse to the *j*th warehouse	km
d_{jk}	Transportation distance from the *j*th warehouse to the *k*th customer	km
Y_j	0-1 decision variable, $Y_j = 1$ when the firm decides to build a forward warehouse at point *j*, and 0 otherwise	
w	Fixed cost of building the distribution outlet of the forward warehouse	yuan
r	Rental cost of building the warehouse of the forward warehouse	yuan
h	Construction cost of building the store for the forward warehouse	yuan
I_j	Alternative point *j* warehouse is required to build a warehouse-store front warehouse, I_j is 1, otherwise 0	
T_j	Overhead cost of a single staff member at store *j* of the warehouse-store model of the front-loading warehouse outlet *j*	yuan
η	denotes the ratio of the coefficient of the operating cost of the front warehouse outlet to the handling volume of fresh products	
θ	Chemical reaction rate, i.e., the damage rate of goods	
t_{ij}	denotes the transportation time from warehouse *i* to warehouse *j*	h
t_{jk}	denotes the transportation time from warehouse *j* to customer *k*	h
ε_1	Cost per unit of cargo damage when transporting from large warehouse *i* to forward warehouse *j*	yuan
ε_2	Unit cost of cargo damage when transporting from warehouse *j* to customer demand point *k*	yuan
$f(t_{ij})$	Time penalty cost	yuan
u_{jk}	Pre-warehouse *j* provides distribution goods to the *k*th customer cluster point	
∂	Consumer risk-bearing cost of building front warehouse *j* of size *v* to distribute goods of unit quality	yuan
ε_2^0	Unit cost of goods loss during transportation in the integrated warehouse-store model	yuan/kg
ε_2^1	Cost per unit of goods lost in transit in the dark warehouse model	yuan/kg
$\propto$	infinitely large	
t_0	Upper time limit for the highest level of customer satisfaction	h
t_f	Lower bound of dissatisfaction indicating that customers cannot tolerate the waiting time	h

penalties for violating time windows, constructing a total cost minimization model for the selection of fresh agricultural product distribution centers. Bo et al. [24] targeted customer satisfaction and built an LIRP optimization model under the satellite warehouse mode to minimize the overall cost of the cold chain logistics system in chain supermarkets.

From the above literature, it can be seen that current research only focuses on constructing models for cold chain distribution and has not fully utilized open-source cold chain logistics big data. Moreover, the cold chain logistics distribution models do not reflect the advantages of last-mile delivery. Therefore, this paper crawls customer demand POI data, performs clustering, establishes suitable warehouse types for clustering center points, and finally considers transportation costs, construction costs, product damage costs, and customer satisfaction from the perspective of enterprises, constructing a multi-objective programming model for pre-warehouse site selection. NSGA-II (multi-objective genetic algorithm) is employed for solving the model to obtain the optimal site selection for pre-warehouses. By determining the optimal solution, the number and locations of pre-warehouses are established, thereby maximizing investment returns in the spatial layout of pre-warehouses.

The highlight of this paper is to compare the effect of two clustering algorithms for dealing with the feature data set of the front warehouse layout, clustering the best customer demand points; comparing the heterogeneous ability of machine learning and integrated learning processing, determining the establishment of the type of warehouses suitable for the clustering of the centre point; constructing a multi-objective planning model established by the selection of the site of front warehouses and solving the optimal set of solutions of the Pareto with the NSGA-II, and, according to the enterprise's own needs, to Select the layout strategy suitable for itself and make scientific decisions.

III. FRONT WAREHOUSE SITING MODEL CONSTRUCTION

A. PROBLEM DESCRIPTION

The research object of this paper is the front warehouse siting problem, and the problem can be described as the logistics network end warehouse layout problem, which needs to take into account the customer demand, the front warehouse, and the city warehouse of the three-level logistics network description. In this paper, through two clustering analysis will be dispersed in the demand area of customer demand points for clustering, clustering results in the center of the cluster on behalf of the point near the customer consumption demand is the most vigorous, the center of each cluster of clusters represent the location of the new customer cluster point, the demand for each cluster center can be predicted through historical data; at the same time, using machine learning algorithms for the selection of the location of the type of pre-positioning warehouse built to classify and determine Establish the type of warehouse suitable for the cluster center point, after determining the type of warehouse, from the perspective of the enterprise to consider the transportation cost, construction cost, operating costs and cargo damage costs and take into account customer satisfaction and service quality, in the limited investment funds and a certain degree of customer satisfaction under the conditions of the optimal location of the front warehouse of the enterprise site to minimize the cost of the largest customer satisfaction program. Through the optimal program to determine the number and location of the warehouse, so that enterprises in the warehouse facilities layout to obtain the maximum return on investment and industry competitiveness.

B. MODEL ASSUMPTIONS

In order to fit the actual situation of the research problem, in the establishment of the mathematical model in this paper, the following conditions are assumed:

(1) The site selection process does not take into account the impact of weather and climate, while not taking into account the problem of competition in the industry;

(2) The goods in the front warehouse are all distributed by the upper level of the city warehouse, and the flow of fresh products at all levels of outlets can not exceed its own maximum inventory;

(3) There is no overloading problem for vehicles in each distribution chain (including vans and riding electric vehicles);

(4) Assuming that the temperature of the fresh products remains constant during the distribution process, the quality of the fresh products in terms of cargo damage is only related to its distribution time, without considering other objective conditions;

(5) The demand for fresh products at each customer cluster point remains constant for a certain period of time, is a constant, and the demand at the demand point within the customer cluster does not exceed the maximum distribution service capacity of the front warehouse;

(6) The large warehouse in the city provides distribution services only for the warehouse-store model front warehouse, and the dark warehouse model front warehouse is distributed only by the warehouse-store model front warehouse, and the transfer of goods between the dark warehouse model front warehouses is not considered;

C. PARAMETER VARIABLES

The meanings of the variables of the basic model are shown in Table 1.

D. MULTI-OBJECTIVE OPTIMIZATION FUNCTION AND CONSTRAINTS

The multi-objective optimization model for the location of forward warehouse outlets is shown in Equation (1), Equation (6) is the objective function of minimizing the cost of the enterprise, and Equation (10) indicates the maximum customer satisfaction.

$$\min Z = \min(F, T) \tag{1}$$

$$C_1 = \left\{ \left(\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} d_{ij} + \sum_{j=1}^{n} \sum_{k=1}^{q} c_{jk} y_{jk} d_{jk} \right) \times 365 \right\} /10000 \tag{2}$$

$$C_2 = \sum_{j=1}^{n} Y_j w = \sum_{j=1}^{n} Y_j \left(I_j h + r \right) \tag{3}$$

$$C_3 = \sum_{j=1}^{n} \sum_{k=1}^{q} Y_j \left(\eta \sqrt{y_{jk}} + I_j T_j \frac{y_{jk}}{365 * 150} \times 12 \right) /10000 \tag{4}$$

$$C_4 = \varepsilon_1 \omega_t + \varepsilon_2 \omega_t = \varepsilon_1 \sum_{m}^{i=1} \sum_{n}^{j=1} x_{ij} \left(1 - e^{-\theta t_{ij}} \right) + \varepsilon_2 \sum_{n}^{j=1} \sum_{q}^{k=1} y_{jk} \left(1 - e^{-\theta t_{jk}} \right) \tag{5}$$

$$\min F = C_1 + C_2 + C_3 + C_4 \tag{6}$$

$$C_5 = \sum_{N}^{j=1} \sum_{q}^{k=1} f\left(t_{jk} \right) u_{jk} \tag{7}$$

$$f\left(t_{ij} \right) = \begin{cases} 0 & 0 \leq t_{jk} \leq t_0 \\ \left[\mu_j \left(t_{jk} - t_0 \right) \right]^{\lambda_j} & t_0 \leq t_{jk} < t_f \\ \propto & t_{jk} \geq t_f \end{cases} \tag{8}$$

$$C_6 = \sum_{N}^{j=1} \sum_{q}^{k=1} \partial y_{jk} u_{jk} \tag{9}$$

$$\min T = C5 + C6 \tag{10}$$

Equation (2) represents the total transportation cost, which includes the transportation cost from the main warehouse to the pre-positioning warehouse and the delivery cost from the pre-positioning warehouse to the customers. The transportation cost is calculated on an annual basis, and the unit of calculation is in ten thousand RMB.

Equation (3) represents the construction cost. If the selected location is for a pre-positioning warehouse in the dark warehouse model, only the rental cost of the warehouse, denoted as "r", needs to be paid. However, if the selected location is for a warehouse-store integrated model, additional costs such as scale rental fees and store decoration costs, denoted as "h", need to be paid.

Equation (4) represents the operational cost, assuming that a store staff member can manage 150 kg of fresh products per day.

Equation (5) represents the cost of goods damage during transportation.

Equation (7) represents the customer time penalty cost function. Based on the fitting test conducted by Liu Xiangtao [25] on four penalty functions, it is known that the concave-convex penalty cost function aligns more with the actual characteristics of customer waiting time in the pre-positioning warehouse. Therefore, this paper chooses the concave-convex time penalty cost function, as presented in Equation (8), to construct the customer satisfaction model based on delivery time.

Equation (9) represents the customer risk cost function. In this paper, the customer risk cost is chosen to represent the impact of service quality on customer satisfaction.

The constraints are as follows:

(1) Binary constraint

All decision variables can only take the values of 0 or 1. Equations (11)–(13), as shown at the bottom of the page.

(2) Distribution capacity constraints

Equation (14) indicates that a customer cluster point can only be served by a forward warehouse, and multiple forward warehouses are not considered to serve a customer demand point;

$$\sum_{j=1}^{\mathrm{n}} u_{jk} = 1 \tag{14}$$

Equation (15) indicates that the total amount of goods from the city warehouse to each front warehouse outlet should be equal to the total amount of distribution from the front warehouse to each customer demand point;

$$\sum_{j=1}^{\mathrm{n}} \sum_{k=1}^{q} y_{jk} = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} \tag{15}$$

Equation (16) indicates that the total amount of transportation from the city warehouse to each forward warehouse is within the maximum supply of the city warehouse;

$$\sum_{j=1}^{\mathrm{n}} x_{ij} \leq R_i \tag{16}$$

Equation (17) indicates that the sum of the amount of goods transported from the city warehouse to the forward warehouse is less than or equal to the maximum capacity of that forward warehouse;

$$\sum_{i=1}^{\mathrm{m}} x_{ij} \leq V_j \tag{17}$$

Equation (18) indicates that the sum of the amount of goods transported from the city warehouse to the forward warehouse is less than or equal to the maximum capacity of that forward warehouse;

$$\sum_{k=1}^{\mathrm{q}} y_{jk} \leq V_j \tag{18}$$

Equation (19) indicates that at least one forward warehouse is laid out in the logistics network;

$$\sum_{j=1}^{\mathrm{n}} Y_j \geq 1 \tag{19}$$

Form (20) that the front warehouse allocation to the customer distribution volume is equal to the customer's

$$Y_j = \begin{cases} 1 \cdots \text{Enterprises build front-loading warehouses at point j} \\ 0 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{or else;} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{cases} \tag{11}$$

$$I_j = \begin{cases} 1 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{Alternative point j requires the construction of} \\ \text{a forward warehouse in the form of a dark warehouse model} \\ 0 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{Alternative point j requires the construction of} \\ \text{an integrated warehouse with a forward warehouse} \end{cases} \tag{12}$$

$$u_{jk} = \begin{cases} 1 \cdots \text{Front-loading bin j for customer demand point allocation at k} \\ 0 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{or else;} \cdots\cdots\cdots\cdots\cdots\cdots \end{cases} \tag{13}$$

demand;

$$\sum_{j=1}^{\mathrm{n}}\sum_{k=1}^{q} y_{jk} = \sum_{k=1}^{q} D_k \tag{20}$$

(3) Equation rounding and non-negative constraints

Formula (21) to formula (24) that the variables take the value of integers and non-negative numbers;

$$i = 1, 2, 3, \cdots, M \tag{21}$$

$$j = 1, 2, 3, \cdots, N \tag{22}$$

$$k = 1, 2, 3, \cdots, q \tag{23}$$

$$x_{ij}, y_{jk} \geq 0 \tag{24}$$

E. ALGORITHM DESIGN OF FORWARD WAREHOUSE SITING MODEL

For the multi-objective planning model established by the front warehouse siting, NSGA-II (multi-objective genetic algorithm) is selected to solve the problem, and the coding in the non-dominated sorting genetic algorithm is expressed as a method to transform the feasible solution of a certain problem from the solution space to the search space that can be handled by the algorithm. Aiming at the characteristics of decision variables Y_j, I_j and u_{jk} in the multi-objective optimization model, this paper adopts binary coding and real number coding, Y_j and I_j, respectively, are expressed as whether to establish a forward warehouse at the point and to establish which kind of warehouse allocation mode of forward warehouse, and they use the set of coding symbols composed of the binary symbols 0 and 1; I_j in the model by the classification algorithm to give the corresponding prediction, in a strict sense of the meaning of a segmented function is not a decision variable, but in the coding process is still chosen to be represented in binary coding. u_{jk} represents the various customer cluster points served by pre-warehouse alternative points. Taking the real number encoding example (3, 2, 1, 4, 3, 3), pre-warehouse alternative point 3 serves customers 1, 5, and 6, pre-warehouse alternative point 2 serves customer 2, pre-warehouse alternative point 1 serves customer 3, and pre-warehouse alternative point 4 serves customer 4.

1) INITIALIZATION STOCK

Initializing the population is to give the initial solution set of the population based on the encoding method and complete the fast non-dominated sorting. The initialization of the population is affected by the range of values of the variables and constraints, generally generated randomly and the range of values between 50-2000, in order to ensure the diversity of the population this paper chooses the initial size of the population is 2000. at this point, it is completed that each solution is assigned a value equal to the value of the rank of the virtual fitness, denoted by *fitness value=irank*.

2) NON-DOMINATED SORTING AND CROWDING CALCULATION

Undominated sorting is performed for each individual of the initialized population until the sorting of all individuals in the population is completed, and then the crowding degree calculation is performed.

3) SELECTION, CROSSOVER AND MUTATION OPERATIONS

① Selection operation: the binary tournament method is used, i.e., two individuals are selected for comparison, and each individual has the same probability of being selected. The specific operation is to pick two genes to be compared with each other, and if the grades of the non-dominated set are not the same, the gene with higher grade will be selected, and if the grades of the non-dominated set are the same, the gene with larger degree of individual crowding will be selected. By using the tournament method, we can find the genetically optimal and relatively optimal individuals. Repeat the above operation until the population size reaches the original size.

② Crossover operation: that is, according to a certain probability of two chromosomal individuals by exchanging part of the gene pairs, the formation of new chromosomes; chromosome crossover determines the algorithm's ability to search globally, so the algorithm is selected to simulate binary crossover. The crossover probability pc as shown in equation (25) reflects the chromosome crossover probability, if the crossover probability is set too low, the search of the algorithm will enter into a stagnant state, which slows down the evolution of the algorithm but improves the global optimization ability; on the contrary, if the crossover probability is set too high, the chromosomes can be adequately paired with each other, but it will produce bad solutions to make the randomization of the selection.

$$p_{\mathrm{c}} = \begin{cases} p_{\mathrm{c,agv}} + \dfrac{G-i}{G} \cdot \left(p_{\mathrm{c,max}} - p_{\mathrm{c,agv}}\right) & d_j(i) < d_{\mathrm{agv}}(i) \\ p_{\mathrm{c,agv}} & d_j(i) = d_{\mathrm{agv}}(i) \\ p_{\mathrm{c,agv}} - \dfrac{i}{G} \cdot \left(p_{\mathrm{c,agv}} - p_{\mathrm{c,min}}\right) & d_j(i) > d_{\mathrm{agv}}(i) \end{cases} \tag{25}$$

where $p_{\mathrm{c,agv}} = 1/2 * \left(p_{\mathrm{c,max}} + p_{\mathrm{c,min}}\right)$ is the average crossover probability, $p_{\mathrm{c,max}}$ denotes the maximum crossover probability, and $p_{\mathrm{c,min}}$ denotes the minimum crossover probability. G is the total number of iterations; i is the current number of iterations; $d_j(i)$ denotes the crowding distance of the jth individual of the current population of the ith generation, and $d_{\mathrm{agv}}(i)$ denotes the average crowding distance of the current population of the i generation. Individuals whose current crowding distance is smaller than the average crowding distance have lower crossover and mutation probabilities, and vice versa.

③ Mutation operation: the probability of mutation p_m shown in equation (26) reflects the probability of the execution of the mutation operation, the smaller the probability of mutation, the higher the stability of the algorithm but is prone to local optimization; the larger the probability of mutation, the population diversity will increase but may destroy the excellent pattern of the optimization search process. This paper is the real number coding used.

$$p_{\mathrm{m}} = \begin{cases} p_{\mathrm{m,agv}} + \dfrac{G-i}{G} \cdot \left(p_{\mathrm{m,max}} - p_{\mathrm{m,agv}}\right) & d_j(i) < d_{\mathrm{agv}}(i) \\ p_{\mathrm{m,agv}} & d_j(i) = d_{\mathrm{agv}}(i) \\ p_{\mathrm{m,agv}} - \dfrac{i}{G} \cdot \left(p_{\mathrm{m,agv}} - p_{\mathrm{m,min}}\right) & d_i(i) > d_{\mathrm{agv}}(i) \end{cases} \tag{26}$$

4) JUDGMENT CONDITIONS FOR ALGORITHM TERMINATION

When judging whether the algorithm ends its operation, it is necessary to stipulate a certain threshold in advance, and if the individual fitness reaches the set threshold, the algorithm ends its operation; or set the maximum number of iterations of the algorithm, which has been converged within the number of iterations can be used as an algorithm termination condition.

IV. FRESH FRONT WAREHOUSE DISTRIBUTION POINT SELECTION

A. THE FIRST CLUSTERING ESTABLISHES THE CLUSTERING AREA OF THE REQUIRED FRONT WAREHOUSE

Outlier detection is an essential part of data mining using Python. Outlier detection is done by discovering features that are significantly different from most objects. The crawled data is clustered by the DBSCAN algorithm, and small clusters far from other clusters are removed according to the clustering results, this process is equivalent to a data cleaning with conditions, and the elimination of outliers can improve the accuracy and cohesion of the secondary clustering, and the flow of the DBSCAN algorithm is shown in Figure 3:

Input: contains n customer demand points x_i dataset $D = (x_1, x_2, \cdots, x_n)$, determine the parameter domain and density threshold Minpts;

Output: the resulting C clusters that satisfy customer needs after clustering analysis;

Step1: arbitrarily select a data point P from the customer demand point dataset D, and test the number of data points K contained in the domain set R of the point P;

Step2: determine the relationship between the number of data points K in the neighborhood and the density threshold Minpts, if K is greater than or equal to Minpts, then data point P is a core point, otherwise data point P is a noise point;

Step3: Perform to find all the customer demand points whose density is reachable and give the label of the class;

Step4: check whether there are unprocessed objects in the dataset in the domain of point P. If there are, continue to perform Step1 and Step2 operations;

Step5: Until all the demand points are labeled as processed and all the customer demand points are categorized into some cluster or as noise points, the algorithm ends the execution.

B. SECOND CLUSTERING TO ESTABLISH CUSTOMER CLUSTER POINTS

1) K-MEANS ALGORITHM

The K-means algorithm, also known as the K-means algorithm, is mainly used to determine the similarity between sample points by calculating the size of the distance between the sample points, so as to classify the points with similar distances in the samples into the same cluster. The core idea of the algorithm is to make the intra-cluster distance as small as possible and the inter-cluster distance as large as possible.

(1) Determination of the number of clustering clusters (K value)

1) Sum of Squared Errors (SSE)

Sum of the Squared Error (SSE), also known as and variance, is calculated by fitting the data and the original data corresponding to the point of the error sum of squares, the size of the function expressed by the fit is good or bad. Generally the smaller the sum of squared error represents the better the clustering effect and the higher the accuracy of data prediction. Its calculation formula is shown in (27):

$$S_E = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i\right)^2 \tag{27}$$

2) Silhouette Analysis (SA)

Silhouette Analysis is a sample assessment using both aggregation and dispersion metrics, which are calculated as shown in (28):

$$s = \frac{b-a}{\max(a, b)} \tag{28}$$

where a represents the similarity between the sample and other sample points within the cluster in which it is located, which is equal to the average distance between the sample point and other sample points in the same cluster, and b represents the similarity between the sample and other clusters, which is equal to the average distance between the sample and all sample points in the nearest cluster. We hope that the clustering results of intra-cluster differences the smaller the better, the greater the differences between the clusters the better, so we hope that b is much larger than a. Profile coefficient s takes the value of the range of [−1,1], and s tends to converge to 1 the better the clustering effect.

(2) K-means algorithm process

In this paper, based on the K-means algorithm for the discrete distribution of customer demand clustering integration to form a strong demand for customer cluster points, the steps are shown below, and its algorithm flow is shown in Figure 4:

Input: data set containing n customer demand points x_i, and the number of clustering clusters K. Output: K clustering clusters that satisfy the clustering requirements.

Step1: Select K points from the sample point points as the center points of the clusters;

Step2: Calculate the distance between sample points, each sample point is divided into the cluster closest to

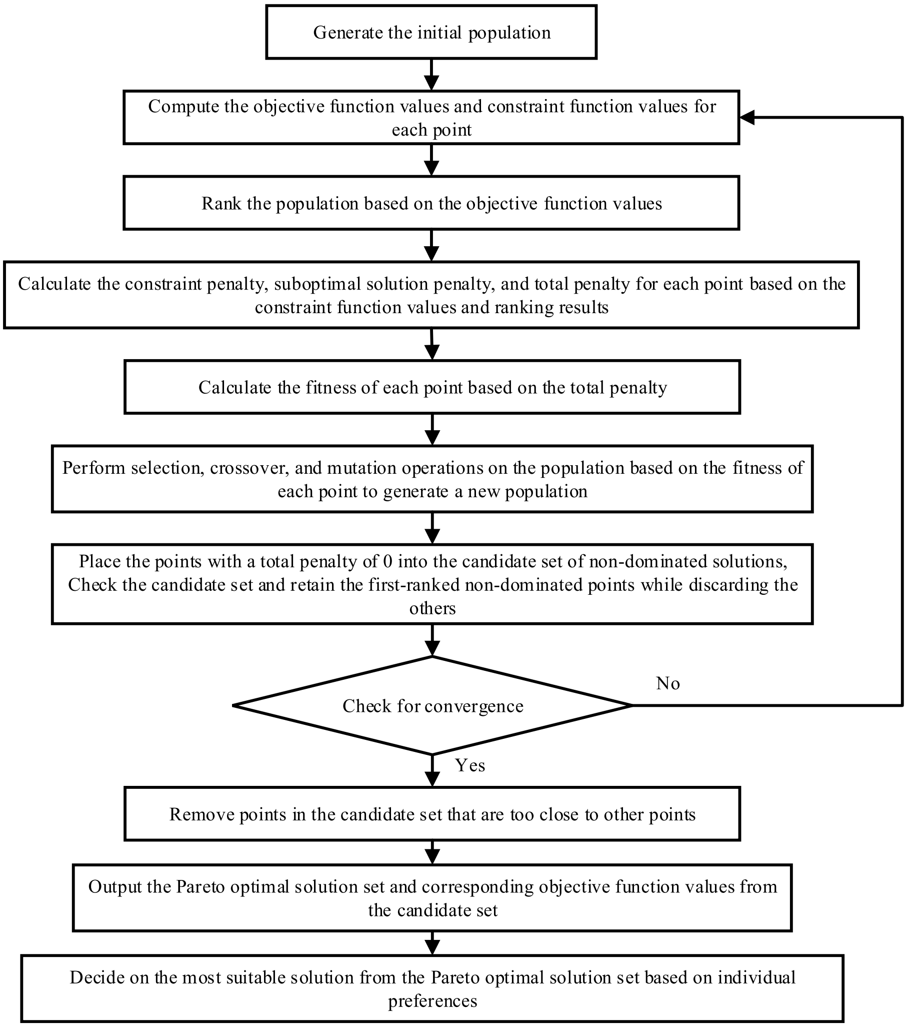

Generate the initial population
Compute the objective function values and constraint function values for each point
Rank the population based on the objective function values
Calculate the constraint penalty, suboptimal solution penalty, and total penalty for each point based on the constraint function values and ranking results
Calculate the fitness of each point based on the total penalty
Perform selection, crossover, and mutation operations on the population based on the fitness of each point to generate a new population
Place the points with a total penalty of 0 into the candidate set of non-dominated solutions, Check the candidate set and retain the first-ranked non-dominated points while discarding the others
Check for convergence
No
Yes
Remove points in the candidate set that are too close to other points
Output the Pareto optimal solution set and corresponding objective function values from the candidate set
Decide on the most suitable solution from the Pareto optimal solution set based on individual preferences

FIGURE 2. Multi-objective genetic algorithm flow chart.

the cluster center, as shown in equation (29), while there are many distance strategies for measuring sample points, generally choose Euclidean distance; because the attributes of the front position in this paper are latitude and longitude coordinates, it is not possible to directly carry out the similarity metric, so we use the formula for calculating the distance between two points by using the latitude and longitude distance formula, and use the formula as the similarity measure formula, the latitude and longitude similarity measure formula is shown in (30):

$$\text{label}_i = \arg\min \parallel x_1 - u_j \parallel \ (\text{where } 1 \leq j \leq k) \tag{29}$$

$$d = 2\arcsin\sqrt{\sin^2\frac{a}{2} + \cos(\text{lat1}) \times \cos(\text{lat2}) \times \sin^2\frac{b}{2}} \times 6378.137 \tag{30}$$

Step3: For the K species clusters generated after clustering, the mean value of the coordinates of the sample points within the cluster is calculated as the center of mass respectively, and the calculation formula is shown in (31):

$$a_j = \frac{1}{|c_i|}\sum_{x \in c_i} x \tag{31}$$

Step4: When the center of mass is not changing or the maximum number of iterations is satisfied, the clustering algorithm stops running.

2) MEAN-SHIFT ALGORITHM

Assuming that the dataset is clustered and analyzed in a multidimensional space, the flowchart of the Mean-shift algorithm is shown in Figure 5.

Step1: Arbitrarily select a point as the center point x from the unclustered data points;

Step2: Find all the points within a bandwidth (i.e., radius h) from the center point x, denoted as S_h, and identify these data points as c clusters;

Step3: taking x as the center point, compute the vector from the center point to each element in the set S_h, and sum the

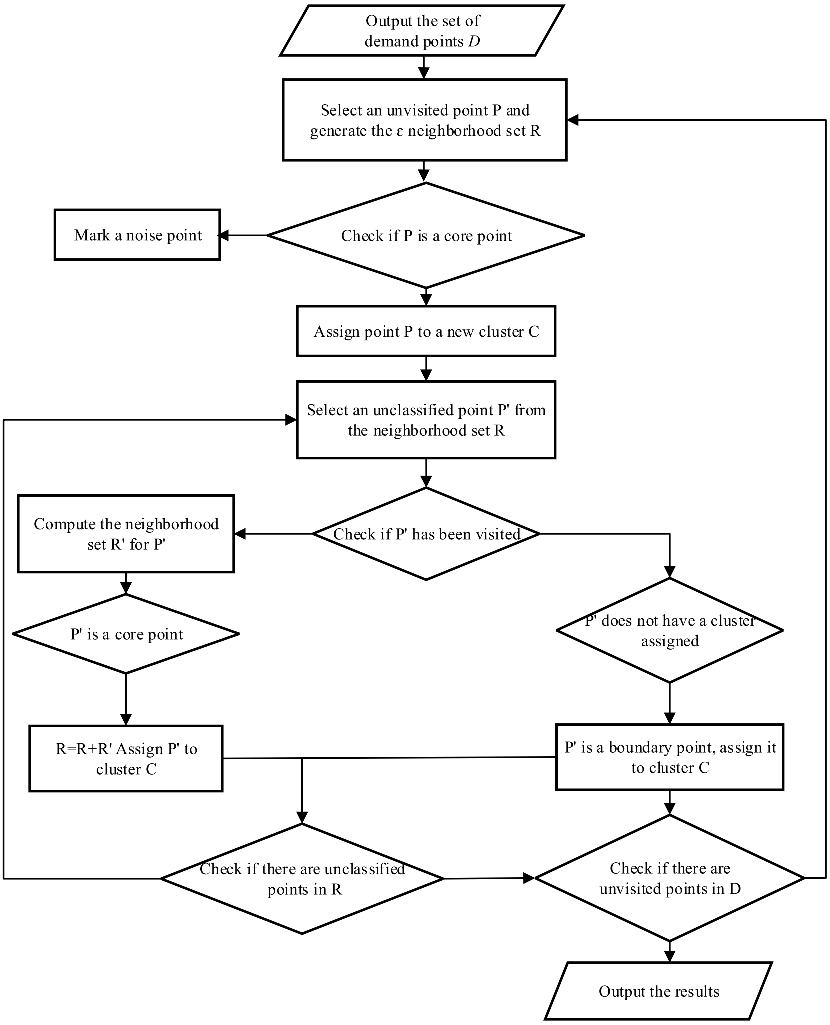

Output the set of demand points D
Select an unvisited point P and generate the ε neighborhood set R
Mark a noise point
Check if P is a core point
Assign point P to a new cluster C
Select an unclassified point P' from the neighborhood set R
Compute the neighborhood set R' for P'
Check if P' has been visited
P' is a core point
P' does not have a cluster assigned
R=R+R' Assign P' to cluster C
P' is a boundary point, assign it to cluster C
Check if there are unclassified points in R
Check if there are unvisited points in D
Output the results

FIGURE 3. Flow chart of DBSCAN algorithm.

vectors by combining the Gaussian kernel function to derive the drift vector $M_h(x)$.

Step4: the center point x moves along the direction of the drift vector $M_h(x)$, and the moving distance is the mode of the offset vector $M_h(x)$;

Step5: Repeat the above steps 2, 3 and 4 until $M_h(x)$ is less than the set threshold demand, in this iteration all points are grouped into cluster c;

Step6: Repeat the above 1, 2, 3, 4, 5 until all data points are labeled;

Step7: When each point is labeled, finalize the cluster to which each point belongs based on the frequency with which the point is accessed;

C. CLASSIFICATION ALGORITHM TO ESTABLISH THE SELECTION OF FRONT WAREHOUSE ALLOCATION POINTS

1) NAIVE BAYESIAN ALGORITHM

The Naive Bayesian algorithm is a classification method based on Bayes' theorem and the assumption of conditional independence of features. Assume that our classification samples are:

$$\left(x_1^{(1)}, x_2^{(1)}, \ldots x_n^{(1)}, y_1\right), \left(x_1^{(2)}, x_2^{(2)}, \ldots x_n^{(2)}, y_2\right), \ldots \left(x_1^{(m)}, x_2^{(m)}, \ldots x_n^{(m)}, y_m\right) \tag{32}$$

There are m samples in the initial data, each sample has n feature indicators, and the output categories of feature indicators are $C_1, C_2, \ldots C_K$ from the initial sample data set can be learned to get the a priori probability distribution $P(X = x)(Y = C_k) = P(X_1 = x_1, X_2 = x_2, \ldots\ldots X_n = x_n \mid Y = C_k)$, then according to the conditional probability formula can be known P, and then using the full probability formula can get the joint probability distribution of X and Y, the joint probability of the formula (33) as:

$$\begin{aligned} P(X = x)(Y = C_k) &= P(X_1 = x_1, X_2 = x_2, \ldots\ldots X_n = x_n \mid Y = C_k) \\ &= P\left(Y = C_k \mid X = X^{(\text{test})}\right) \end{aligned} \tag{33}$$

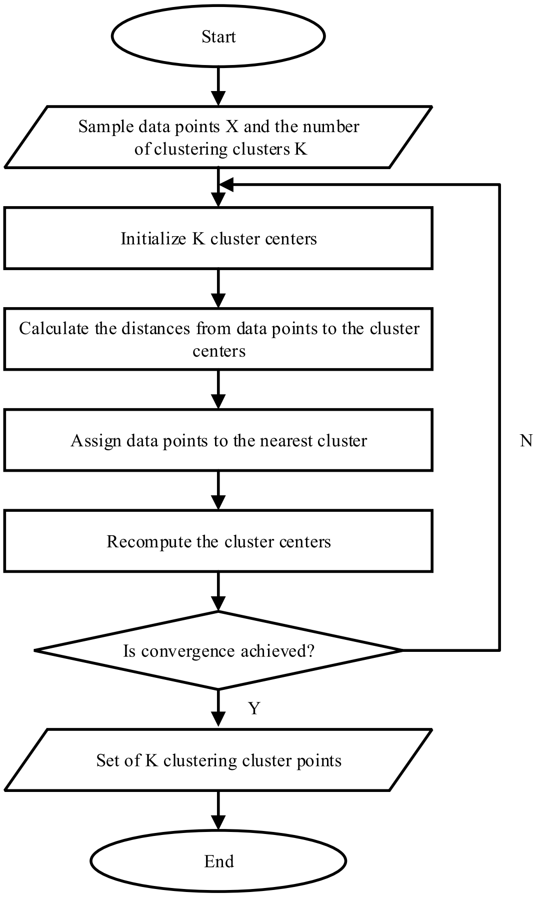

Start
Sample data points X and the number of clustering clusters K
Initialize K cluster centers
Calculate the distances from data points to the cluster centers
Assign data points to the nearest cluster
Recompute the cluster centers
Is convergence achieved?
N
Y
Set of K clustering cluster points
End

FIGURE 4. **Flow chart of K-means algorithm.**

Thus it is only necessary to compute $P\left(Y=C_k \mid X=X^{(\text{test})}\right)$ under all k conditional probabilities and then find the category corresponding to the maximum conditional probability to complete the classification.

2) XGBOOST ALGORITHM

XGBoost algorithm is also one of the integrated learning algorithms, which internally implements the Gradient Boosted Tree (GBDT) model and improves and optimizes the algorithms in the model, which can be used in regression problems as well as data classification problems.The principle of the XGBoost algorithm and the formulas are pushed to the following:

1) Determine the objective function in the algorithm:

The objective function in the XGBoost algorithm is the sum of the loss value and the regular expression, assuming that the training set $D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$, $(|D|=n, X_i\in R^m, y\in R)$ samples. The algorithm base model for lifting the decision tree, as shown in equation (34):

$$f_M(x)=\sum_{M}^{m=1} T(x;\theta_m) \tag{34}$$

where $T(x;\theta_m)$ denotes a decision tree, θ_m denotes is the parameter of the decision tree, and M is the number of decision trees. Then again, according to the forward distribution algorithm, it is known that the initial boosted decision tree $f_0(x)=0$. After m iterations, the model is $f_m(x)=f_{m-1}(x)+T(x;\theta_m)$, where $f_{m-1}(x)$ represents the accumulation of the previous m-1 trees, which is a constant.

The objective function of XGBoost is shown in (35):

$$Obj^{(t)}=\sum_{i=1}^{n} L\left(y_i,\hat{y}^{(t-1)}+f_t(x_i)\right)+\Omega(f_t)+\text{ constant} \tag{35}$$

where: $\sum_{i=1}^{n} L\left(y_i,\hat{y}^{(t-1)}+f_t(x_i)\right)$ is the loss function, which represents the difference between the true value and the predicted value; $\Omega(f_t)$ is the regularization term, which is an optimization made by XGBoost in the traditional GDBT, as shown in equation (36):

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda \parallel w\|^2 \tag{36}$$

where T is the number of leaf nodes, $\|$ w $\|$ is the mode of the leaf node vector, and λ and r are two hyper parameters used to regulate the degree of influence of the decision tree.

2) The loss function is approximated and represented using the second-order Taylor expansion:

The second order Taylor expansion is shown in equation (37):

$$f(x_0+\Delta dx)\approx f(x_0)+f'(x_0)\Delta x+\frac{f''(x_0)}{2}(\Delta x)^2 \tag{37}$$

A second-order Taylor expansion of $f_t(x_i)$ in the loss function shows that:

$$\begin{aligned}\text{Obj}^{(t)}\approx&\sum_{i=1}^{n}\left[L\left(y_i,\hat{y}^{(t-1)}\right)+g_i*f_i(x_i)+\frac{1}{2}h_j*f_i(x_i)^2\right]\\&+\Omega(f_i)+\text{constant}\end{aligned} \tag{38}$$

where g_i and h_i expressions are (39), (40) and both are constants:

$$g_i=\partial_{\bar{y}^{(t-1)}}L\left(y_i,\hat{y}^{(t-1)}\right) \tag{39}$$

$$h_i=\partial^2_{\bar{y}^{(i-1)}}L\left(y_i,\hat{y}^{(t-1)}\right) \tag{40}$$

Again, since $L\left(y_i,\hat{y}^{(t-1)}\right)$ is also a constant, the objective function expands as:

$$\begin{aligned}Obj^{(t)}=&\sum_{i=1}^{n}\left[g_i*f_t(x_i)+\frac{1}{2}h_i*f_t(x_i)^2\right]\\&+\Omega(f_t)+\text{constant}\end{aligned} \tag{41}$$

3) Add the regular expression and simplify to get the objective function expression (42):

$$\begin{aligned}obj^{(t)}=&\sum_{n}^{i=1}\left[g_i*f_i(x_i)+\frac{1}{2}h_i*f_t(x_i)^2\right]+\gamma T+\frac{1}{2}\lambda\sum_{T}^{j=1}w_j^2\\=&\sum_{n}^{i=1}\left[g_i*w_{q(x_i)}+\frac{1}{2}h_i*w_{q(x_i)}^2\right]+\gamma T+\frac{1}{2}\lambda\sum_{T}^{j=1}w_j^2\end{aligned}$$

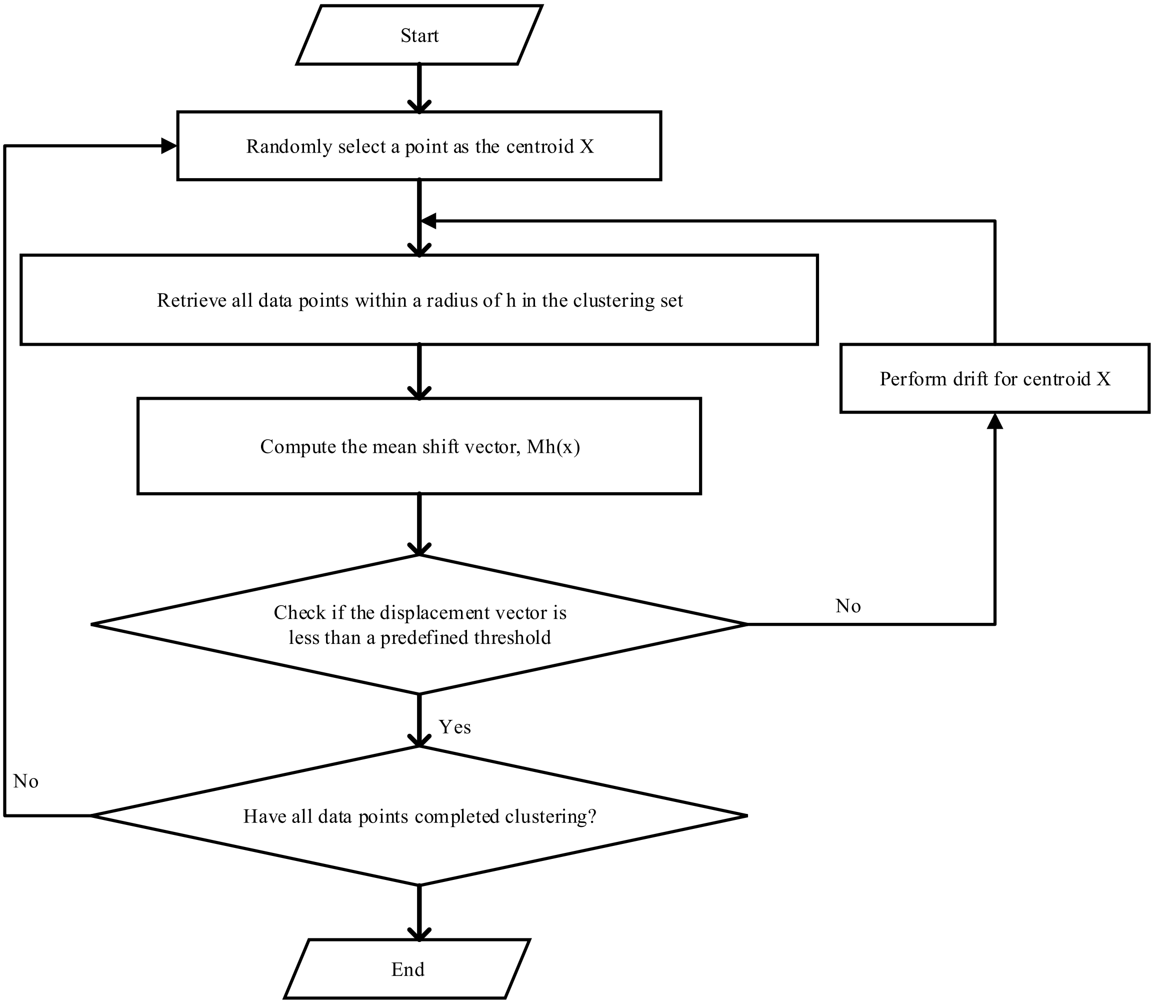

Start
Randomly select a point as the centroid X
Retrieve all data points within a radius of h in the clustering set
Perform drift for centroid X
Compute the mean shift vector, Mh(x)
Check if the displacement vector is less than a predefined threshold
No
Yes
No
Have all data points completed clustering?
End

FIGURE 5. Flow chart of mean-shift algorithm.

$$
\begin{aligned}
&= \sum_{T}^{j=1}\left[G_j w_j + \frac{1}{2}\left(H_j+\lambda\right) w_j^2\right] + \gamma T \\
&= -\frac{1}{2}\sum_{T}^{j=1}\frac{G_j^2}{H_j+\lambda} + \gamma T \qquad (42)
\end{aligned}
$$

V. ANALYSIS OF THE CALCULATIONS

A. DATA ACQUISITION

1) CRAWLING OF POI DATA FOR CUSTOMER DEMANDS

Taking Huangdao District of Qingdao City as the research object, we first query the geographic code of the Gaode map of Huangdao District, including the city code and administrative division code, and then apply for the key of the Web service, and then the name information of each demand point in Huangdao District of Qingdao City, obtained through the Web crawler program, as well as latitude and longitude coordinates are imported into ArcMap 10.7, and then the data is visualized, so that we can get the spatial distribution of the demand points in Huangdao District, as shown in Figure 6. distribution is shown in Figure 6:

The customer demand data points are imported into Arcgis software, and after eliminating the duplicates in the data points obtained by the crawler, a total of 11,539 data points are imported, of which the types of data points include colleges and universities, communities, shopping malls, subway stations, bus stops, entertainment venues, express delivery sites, catering services, and so on. From the spatial distribution map of consumer demand points, it can be seen that consumers mainly appear in the coastal zone of the West Coast Economic Development Zone, where the economy is more developed and the population distribution is more dense, while the population distribution of other townships and streets in Huangdao District is more dispersed compared with here. Therefore, a secondary cluster analysis is chosen for site selection and layout of service points.

2) CRAWLING DATA FOR CLASSIFICATION OF FRONT-LOADED WAREHOUSE DISTRIBUTION PATTERNS

Due to the small number of layouts of Freshippo and MissFresh in Qingdao, it is not possible to choose the

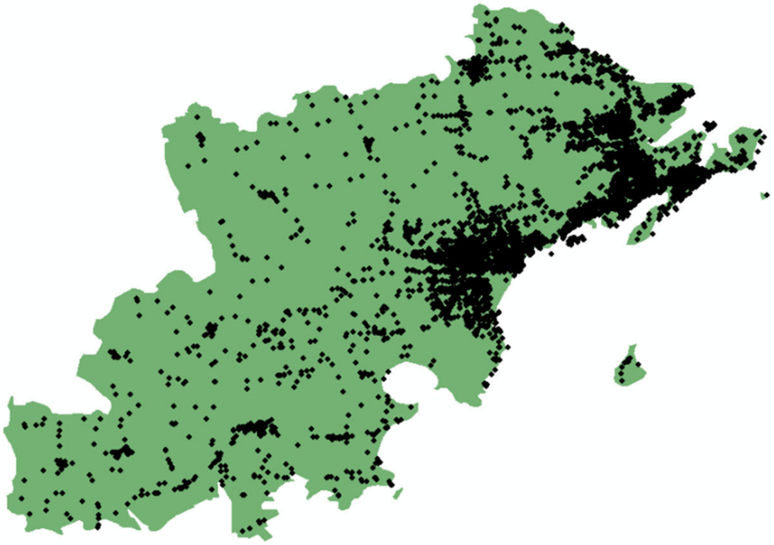


FIGURE 6. Spatial distribution of consumer demand points in Huangdao District.

layout data of front-loaded warehouses in Qingdao as the training set. Therefore, the data of two cities similar to Qingdao in terms of economic development level, population size, and transportation conditions: Nanjing and Wuhan are chosen as the training set. Firstly, the spatial geographic coordinates of Freshippo and MissFresh in Nanjing and Wuhan on the Gaode map are crawled respectively, and then the spatial coordinates of these stores are used as the center point to crawl the feature indicators within 3 km nearby, as shown in Table 2 for some feature indicator data in Wuhan.

B. CLUSTERING RESULTS AND DEMAND FORECASTING

1) RESULTS OF THE FIRST CLUSTERING ANALYSIS

The crawled data are clustered by DBSCAN algorithm, and small clusters far away from other clusters are shaved off according to the clustering results. The POI data of Huangdao District of Qingdao City, Qingdao, crawled from Gaode map is plotted by Matplotli in python to plot the clustering results, as shown in Figure 7, the number of clusters output by DBSCAN algorithm is 4, in which the blue cluster contains 9757 demand points, which is the largest cluster; while green, yellow and red are all smaller clusters, and the black color is the outlying point of the first clustering, and the black color of the outlying points and smaller number of clusters are shaved off, and the remaining 9757 data points are imported into Arcgis for visualization, as shown in Figure 8.

2) RESULTS OF THE SECOND CLUSTERING ANALYSIS

a: ANALYSIS OF MEAN-SHIFT CLUSTERING RESULTS

After the data cleaning of DBSCAN clustering algorithm, Mean-shift algorithm is firstly used to realize the clustering of customer demand points, and the clustering results are shown in Figure 9, Mean-shift algorithm divides 9757 POI data into 42 clusters, and according to the ratio of the data points in each cluster to the total number of customer demand points, the average daily demand of cold chain logistics products in each cluster is obtained. The coordinates of each cluster center point and the number of demand points within the cluster are shown in Table 3:

b: K-MEANS CLUSTERING RESULTS ANALYSIS

The K-means clustering algorithm was used to analyze the POI data points, and the optimal number of clusters and cluster centroids were solved by calling the K-means package in Python and setting the parameters. From the characteristics of the data, it can be seen that the value of K is between 5-15, and the maximum number of iterations is set to 250, the clustering effect is evaluated as shown in Figure 10, when the value of K is taken as 10-15 the clustering effect is better, but when K is taken as 10 the slope of the sum of the squares of the error varies greatly, and the value of the profile coefficients is larger compared to the other values between 10-15, so the final number of clusters is determined to be 10, the final number of clusters is determined to be 10. At this time, the clustering results are shown in Fig. 11, and the coordinates of the clustering midpoint of each cluster and the number of demand points in the cluster are shown in Table 4.

c: COMPARISON OF CLUSTERING RESULTS

In the urban distribution network, the front warehouse outlets are directly facing the customers, with the characteristics of limited service radius and handling volume, and the layout of front warehouses generally needs to be arranged in the places where the customers are concentrated. As shown in Table 2 and 3, the Mean-shift algorithm divides POI data

TABLE 2. Partial data of the training set of Wuhan City for feature indicators.

Designation	Coordinates	Number of residential areas	Number of schools	Number of supermarkets	Number of courier stations	Number of metro stations	Number of bus stops	Number of places of entertainment	Number of companies	Front position type
MissFresh(Hidden Dragon New Town Store)	114.4269,30.4410	174	28	446	66	4	86	277	868	1
MissFresh(South Lake Avenue Store)	114.3170,30.4747	183	33	553	66	11	72	242	788	1
MissFresh(Gutian Store)	114.1972,30.6052	272	31	656	78	6	126	351	859	1
MissFresh(Polytechnic University Store)	114.3358,30.5197	573	87	880	141	18	131	875	852	1
Freshippo (Greenland Central Plaza)	114.1535,30.4860	134	13	305	48	4	104	201	786	0
Freshippo (Luitou Plaza)	114.3210,30.3726	175	17	512	58	3	107	247	607	0
Freshippo (Methodist)	114.2815,30.5850	892	112	870	169	18	196	876	855	0
Freshippo (Jiangcheng Avenue Store)	114.2169,30.5162	136	22	429	81	11	122	252	552	0

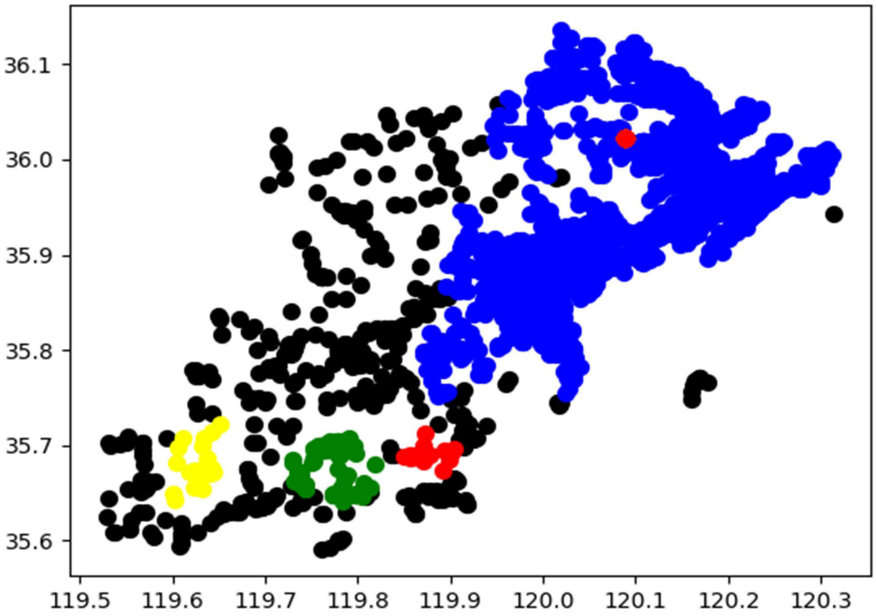


FIGURE 7. The first clustering result graph.

points into 41 clusters with a radius of 3 kilometers, and there are only a few dozens of sample points in clusters 14-19, which is less than 1% of the total samples, and the sample points in clusters 21-38 account for less than 0.5%, and the

TABLE 3. Mean-shift customer cluster point information.

Customer Clustering Points	latitude and longitude coordinates	Number of demand points in the cluster	Percentage of demand points (%)
0	120.1852，35.9606	1966	20.15
1	119.9976，35.8766	1092	11.19
2	120.2190，35.9582	1048	10.74
3	120.1634，35.9364	732	7.50
4	120.0369，35.8848	1118	11.45
5	120.1539，36.0155	670	6.86
⋮	⋮	⋮	⋮
39	119.9188，35.9382	5	0.051
40	119.9609，36.0556	4	0.041
41	119.9336，35.7792	4	0.041

TABLE 4. K-means customer cluster point information.

Customer Clustering Points	latitude and longitude coordinates	Number of demand points in the cluster	Percentage of demand points (%)
C1	120.1439,36.0084	1029	10.546%
C2	119.9872,35.8797	1229	12.596%
C3	120.1783,35.9567	1053	10.792%
C4	120.0079,36.0667	339	3.474%
C5	120.0407,35.8872	2846	29.169%
C6	120.2178,36.0359	583	5.975%
C7	120.2352,35.9639	1349	13.826%
C8	120.1235,36.0808	428	4.387%
C9	119.9876,35.8230	260	2.665%
C10	120.1173,35.9194	641	6.570%

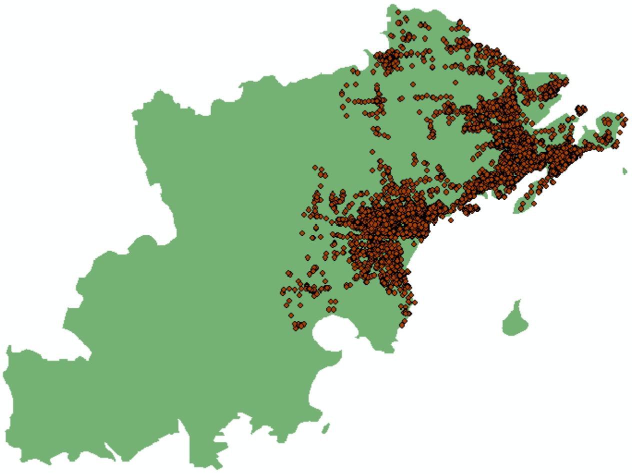


FIGURE 8. Distribution of demand points after elimination of outliers.

sample points in clusters 39-41 account for less than 0.1%, while the sample points in clusters 1-5 account for a higher percentage of the sample points, which leads to a larger gap in point Density will have a large gap, demand point distribution uneven phenomenon; on the contrary, K-means algorithm in the sample point classification uniformity degree has a better performance, sample point less clusters accounted for 2.6%, Mean-shift algorithm is much better than 0.1% of the situation, so choose K-means algorithm clustering results as a new customer customer cluster points to carry out follow-up research.

C. CLASSIFICATION RESULTS AND SELECTION OF WAREHOUSE DISTRIBUTION POINTS

1) ANALYSIS OF CLASSIFICATION RESULTS

For the selection of classification algorithms, we compare the ROC curves and AUC values of the four algorithms, namely, Naive Bayes, Support Vector Machine, Decision Tree and XGBoost, as shown in Figure 12. Through comparison, we found that the AUG area of the XGBoost algorithm is the largest, and the model algorithm result output is more stable, so we choose the XGBoost algorithm for solving the problem of front-loading warehouse allocation mode selection.

The factors considered in the layout of the front warehouse in each city are basically the same, so the data of Nanjing and Wuhan can be used as a training set for classification training. In this paper, we obtained a total of 150 coordinates of Freshippo and MissFresh in Nanjing and Wuhan, and 127 data were obtained through preliminary data cleaning. The 10 customer cluster points obtained by clustering are

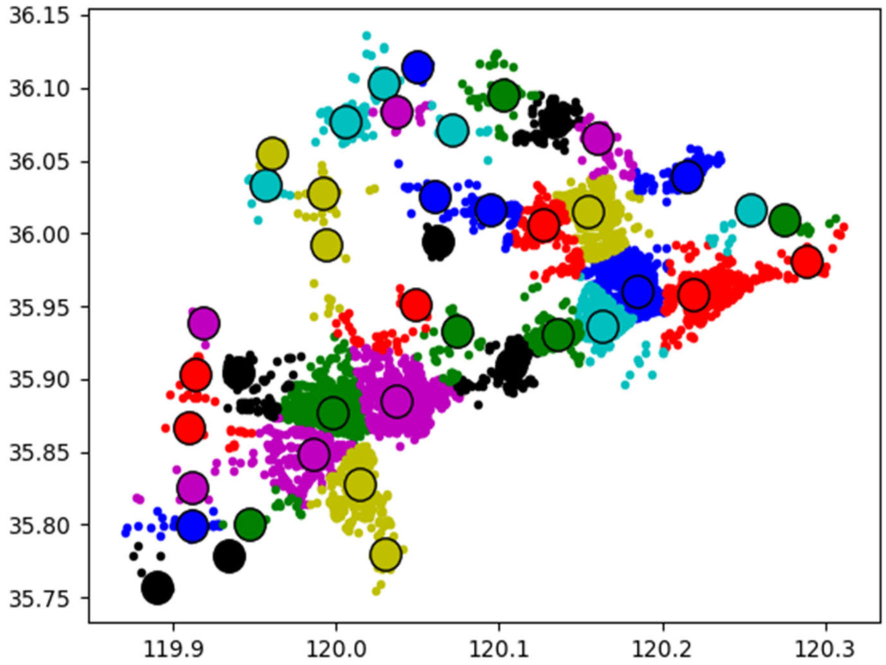


FIGURE 9. The second clustering result graph.

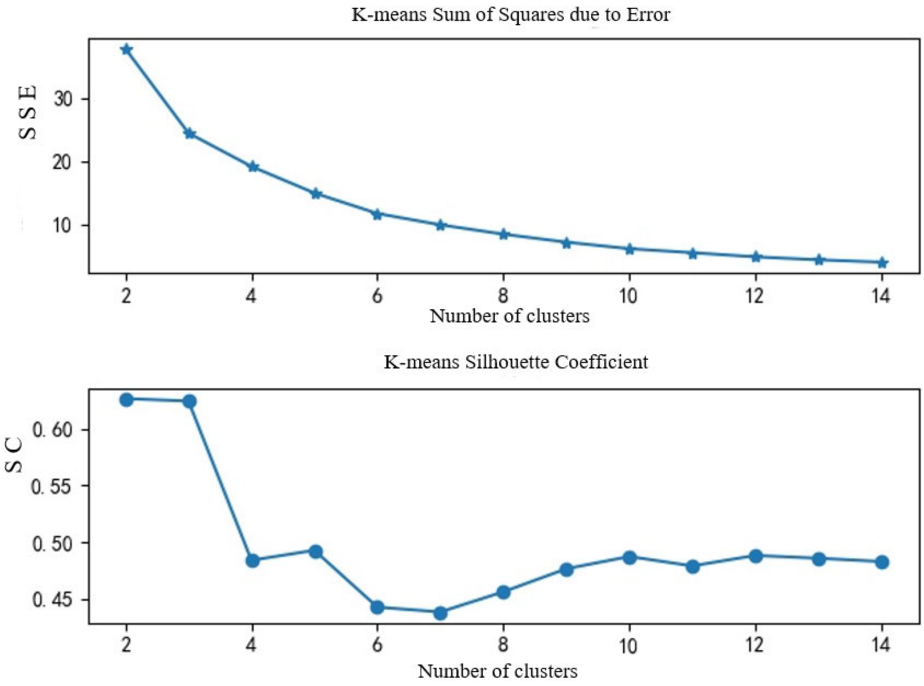


FIGURE 10. K-means clustering evaluation.

input into the model as a test set for prediction, and the results are shown in Table 5:

From the classification results table, it can be seen that customer cluster points C1, C2, C3, C5, and C7 establish front warehouses in warehouse-shop mode, and customer cluster points C4, C6, C8, C9, and C10 establish front warehouses in ‘‘dark warehouse’’ mode.

2) SELECTION OF DELIVERY POINTS FOR FRONT-LOADING WAREHOUSES

Based on the classification results of the machine learning algorithm on the warehouse and distribution modes at the locations of the customer cluster points, the warehouse-store integrated mode with a large floor space and a rich variety of SKUs is taken as an alternative location, and the dark

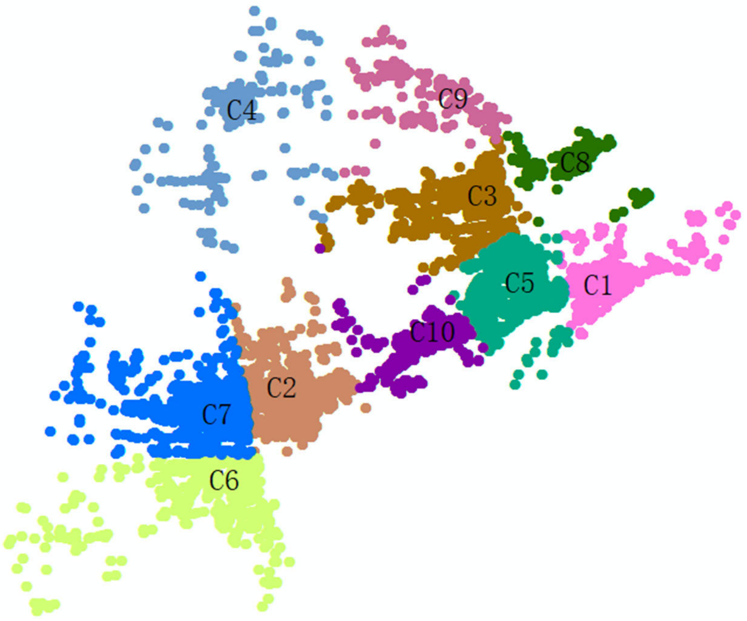
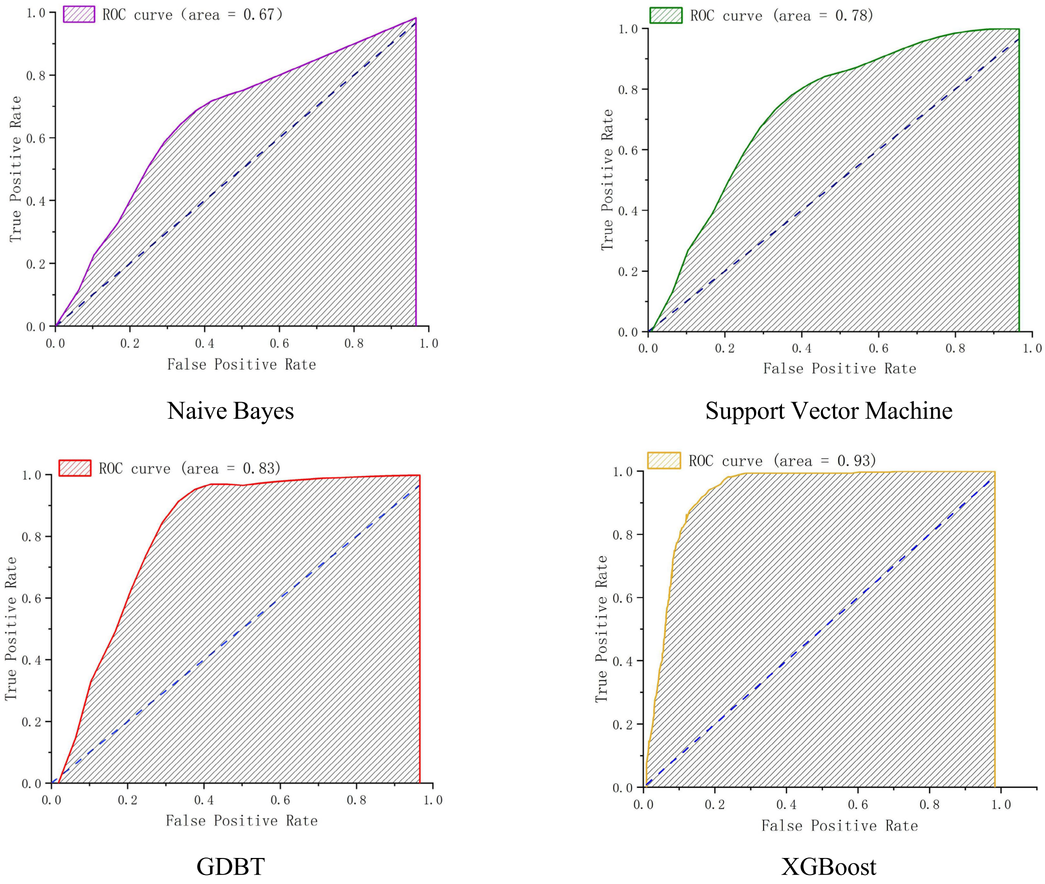


FIGURE 11. K-means clustering result.



Naive Bayes

Support Vector Machine

GDBT

XGBoost

FIGURE 12. ROC graphs and AUC area graphs of the four algorithms.

warehouse mode with a small floor space, a smaller inventory capacity and no offline business activities is taken as a distribution point. That is, customer cluster points C1, C2, C3, C5 and C7 are used as alternative points for warehouse-store

TABLE 5. XGBoost model prediction results display.

Designation	Coordinates	Number of residential areas	Number of schools	Number of supermarkets	Number of courier stations	Number of metro stations	Number of bus stops	Number of places of entertainment	Number of companies	Front position type
C1	120.14396,36.00845	123	12	171	60	3	96	146	173	0
C2	119.98726,35.87974	188	22	169	6	5	119	168	162	0
C3	120.17831,35.95672	182	34	167	88	9	142	166	176	0
C4	120.00800,36.06672	32	5	89	4	0	39	52	173	1
C5	120.04073,35.88719	183	16	177	86	8	141	167	170	0
C6	120.21782,36.03590	47	6	129	17	0	41	73	165	1
C7	120.23524,35.96388	169	16	175	56	4	95	126	155	0
C8	120.12351,36.08082	28	6	72	8	3	50	27	175	1
C9	119.98766,35.82301	58	18	140	26	2	76	35	162	1
C10	120.11728,35.91942	75	13	167	40	6	65	120	163	1

integration front warehouses, with centralized distribution by city warehouses; and customer cluster points C4, C6, C8, C9 and C10 are used to establish front warehouses in the "dark warehouse" mode, with front warehouses in the warehouse-store integration mode of C1, C2, C3, C5 and C7 being used for distribution. Goods. The center point of the cluster is used as the center point for heat map analysis, and a circular area with a radius of 5 kilometers is established. The darker parts of the heat map are extracted as candidate points for the front-loading warehouse outlets, and Figure 13 shows that the darker the color, the more economically developed the region is, and the higher the population density is, the more suitable it is to be an alternative point. Finally, five alternative points that can be used for the construction of front-loading warehouse distribution outlets are identified in the economic region.

D. MODEL SOLUTION ANALYSIS

1) MODEL PARAMETER SETTING

(1) Distance and time of city warehouse, alternative point of front warehouse and customer gathering point

There is only one city warehouse distribution center of R logistics enterprise in Huangdao District, which is located in the west side of Huanghe Road, so the number of city warehouses is 1. Cold-chain transport vehicles are used for distribution service between city warehouses and alternative points, and the transport speed is 50 km/h. The unit transport cost from city warehouses to alternative points of the front warehouses is 0.03 Yuan/(km·kg), and the supply capacity of city warehouses is 3,000,000 kg. The distance from the city warehouse to each alternative point of the forward warehouse is shown in Table 6.

In the transportation process from the front warehouse alternative point to the customer cluster point, the transportation mode used by the distributor is a motorcycle, and the shortest path between the front warehouse alternative point and the customer cluster point is obtained by calling the API of riding planning route to obtain the distributor's distribution route as shown in Table 7, and according to the actual distance and time data of the running distance and time to obtain the distribution speed is 24 km/h, and the unit of the customer cluster point to the front warehouse alternative point Transportation cost is \$0.2/(km·kg), and the shortest time from the customer cluster point to the alternative point is shown in Table 8.

(2) Fixed cost of constructing front warehouses

According to on-site research, it is found that the prepositioning warehouse in R Company's integrated store model generally covers an area of 300-1000 m^2. Based on the online data from the IoT Cloud Warehouse in the Huangdao district, the rental cost for cold chain warehouses is relatively

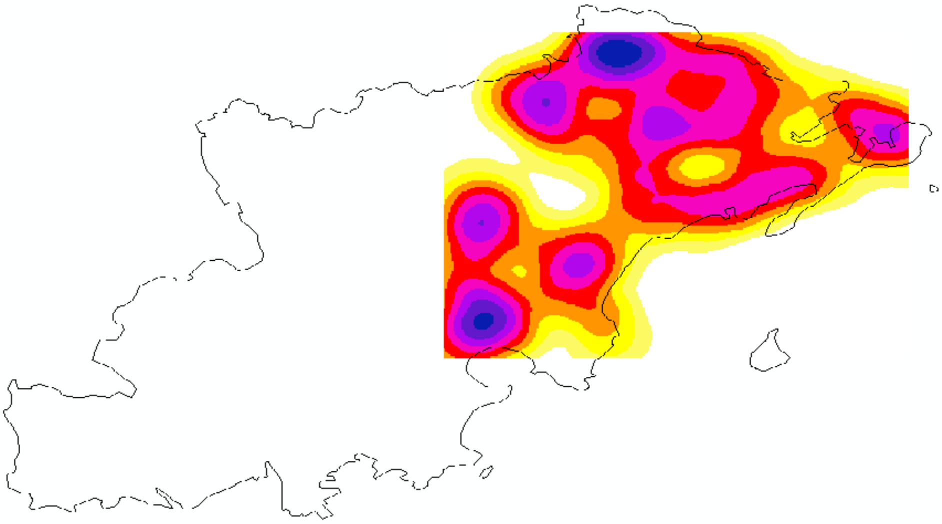


FIGURE 13. Customer cluster point heat map analysis.

TABLE 6. Distance between city warehouse and each alternative point.

Alternative point	Distance (unit/km)	Time (unit/h)
J1（C7）	18.10	0.36
J2（C2）	20.10	0.40
J3（C5）	18.61	0.37
J4（C1）	7.68	0.15
J5（C3）	14.18	0.28

TABLE 7. Distance between customer cluster points to each alternative point.

Distance	JI(unit/km)	J2(unit/km)	J3(unit/km)	J4(unit/km)	J5(unit/km)
Cl	13.53	27.42	21.91	0	8.59
C2	26.27	0	6.06	26.89	22.29
C3	6.01	21.97	16.75	8.30	0
C4	30.62	23.19	24.00	19.09	26.39
C5	20.82	5.76	0	22.00	16.84
C6	15.87	33.95	28.73	9.67	12.90
C7	0	26.20	20.98	13.31	5.99
C8	21.13	33.45	32.72	13.21	16.90
C9	30.48	8.09	10.80	31.65	26.50
C10	12.31	13.91	8.69	13.49	8.33

high, with an average of 45 yuan per square meter per month. The investment capital for leasing a single prepositioning warehouse needs to range from 120,000 to 350,000 yuan. Due to the integration of the store and warehouse model, the prepositioning warehouse requires renovation for the store, which occupies 30% of the prepositioning warehouse area. The renovation cost is 900 yuan per square meter. R Company needs to allocate 100,000 to 270,000 yuan for warehouse renovation and reconstruction expenses. The inventory capacity is between 500 and 800 tons. The construction costs of candidate prepositioning warehouses are summarized in Table 9.

(3) Penalty cost function of waiting time

According to the research results of Liu Xiangtao [25] on the fit of the parameters of the penalty cost function, it was found that the concave-convex type time waiting function curve is more consistent with the actual, combined with the distribution time data of company R, the time sensitivity coefficient μ_j and λ_j are set to 8 and 2, respectively, and the relevant parameters are brought to the function can be derived

TABLE 8. Time between customer cluster points and each alternative point.

Distance	JI(unit/h)	J2(unit/h)	J3(unit/h)	J4(unit/h)	J5(unit/h)
Cl	0.56	1.14	0.91	0.00	0.36
C2	1.09	0.00	0.25	1.12	0.93
C3	0.25	0.92	0.70	0.35	0
C4	1.28	0.97	1.00	0.80	1.10
C5	0.87	0.24	0.00	0.92	0.70
C6	0.66	1.41	1.20	0.40	0.54
C7	0.00	1.09	0.87	0.55	0.25
C8	0.88	1.39	1.36	0.55	0.70
C9	1.27	0.34	0.45	1.32	1.10
C10	0.51	0.58	0.36	0.56	0.35

TABLE 9. Construction cost of each candidate front-end warehouse.

Forebay option point number	Warehouse rental cost (ten thousand yuan per year)	warehouse renovation cost (ten thousand yuan per year	prepositioned warehouse storage capacity (tons)
J1（C7）	20	16	500
J2（C2）	17	14	500
J3（C5）	35	25	800
J4（C1）	17	14	500
J5（C3）	18	15	500

TABLE 10. Other related parameters.

Parameters	Parameter Meaning	Unit	Numerical value
θ	Damage rate of goods		0.005
ε_1	Unit cost of goods damage during transportation from urban warehouse to prepositioned warehouse	yuan/kg	5
ε_2^0	Unit cost of goods damage during transportation in the warehouse-store integration model	yuan/kg	10
ε_2^1	Unit cost of goods damage during transportation in the "dark warehouse" model	yuan/kg	6
$\propto$	Infinity		10000
t_0	Time limit for the highest level of customer satisfaction	h	0.5
t_f	Lower limit of customer dissatisfaction representing the intolerable waiting time	h	1.2
∂	Consumers bear the risk when receiving quality goods from prepositioned warehouses	yuan/kg	0.008

from the formula (43).

$$f\left(t_{jk}\right)=\begin{cases}0 & 0\leq t_{jk}\leq 0.5\\ \left[8\left(t_{jk}-t_0\right)\right]^2 & 0.5\leq t_{jk}<1.2\\ \propto & t_{jk}\geq 1.2\end{cases} \tag{43}$$

(4) Other related parameters

The cargo loss rate of fresh products is between 3% and 10%, and the cargo loss rate of the integrated warehouse-store mode is higher than that of the "dark warehouse" mode, so the cargo loss parameters and other parameters are set as shown in Table 10.

2) SITING SOLUTION RESULTS

The use of MATLAB-R2021b programming to realize the genetic algorithm based on the front warehouse siting model solution, the algorithm solves the iterative curve

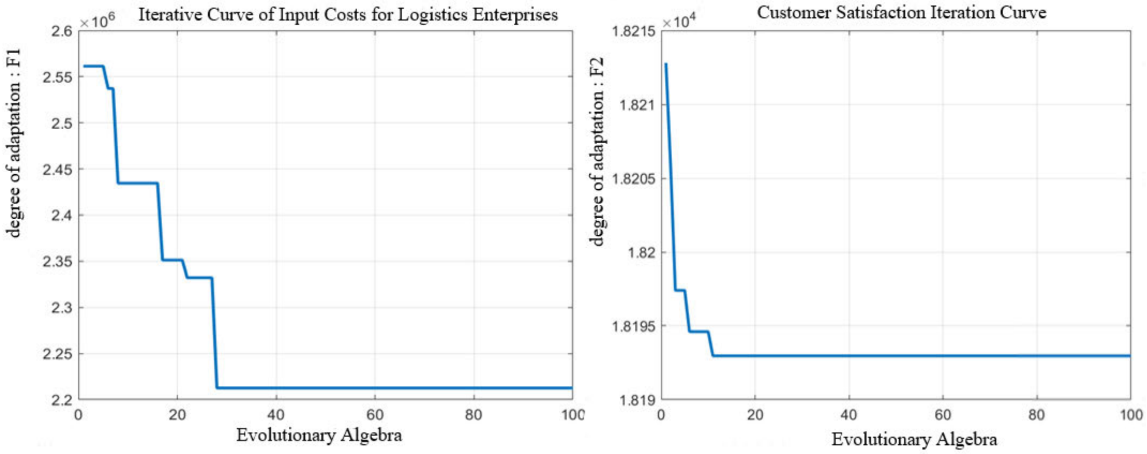


FIGURE 14. Algorithms for solving iterative curves.

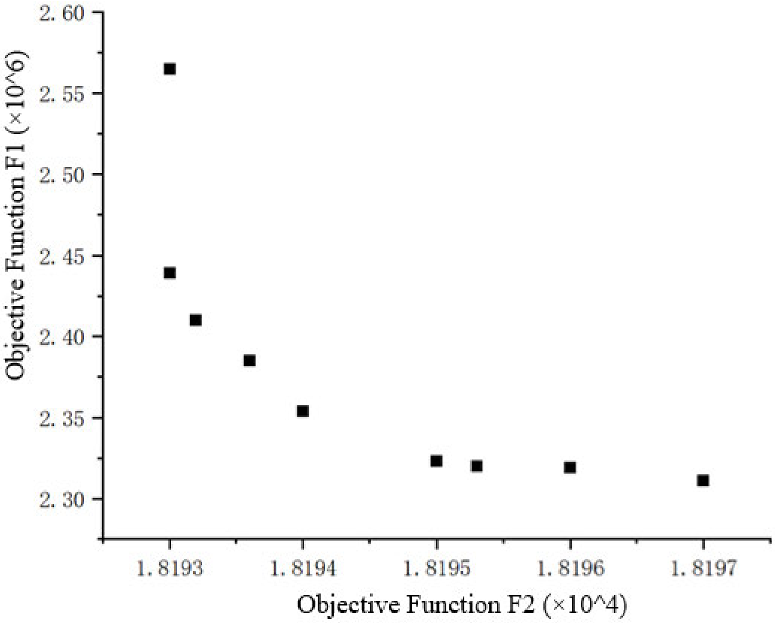


FIGURE 15. Pareto optimal solution set.

shown in Figure 14. In this paper, a total of 5 warehouse store front warehouse preparation, 5 dark warehouse mode front warehouse, according to the model solution obtained to determine the number of front warehouse were 1, 2, 3, 4, 5 for classification and discussion, and then discuss the pareto non-inferiority solution in 5 cases, to determine the final siting plan of the front warehouse.

When $n = 1$, there are 10 options for the selection of candidate front warehouses, i.e., selecting one of the 5 warehouse-store front warehouse alternatives to provide service to the other 9 front warehouse (4 warehouse-store and 5 dark warehouse) alternatives, at which point the constraints Eqs. (17) and (18) cannot be satisfied, and thus no feasible solution can be generated. When $n = 2$ and $n = 3$, consistent with $n = 1$, the constraints cannot be satisfied and thus there is no solution. Front-loading Warehouse When $n = 4$, there are three options for the selection of candidate front-loading warehouses, i.e., selecting four of the five warehouse store one candidate front-loading warehouses to provide service to the remaining demand points. When $n = 5$ the same way, Pareto optimal solution is more concentrated, the gap between the frontier solutions is not large body now the penalty cost is smaller, the enterprise cost investment is larger, the run results to get a set of pareto optimal solution set shown in Figure 15.

Firm R can choose different decision options depending on the needs.

① When company R pays special attention to customer satisfaction and is willing to maintain customer

TABLE 11. Results of site selection of front warehouse.

Solution Results	Objective Function 1	Objective Function 2	Number of Selected Locations	Distribution Plan
Plan 1	1.287136×10^6	6.038×10^4	4	J2 serves k2、k9； J3 serves k5、k7； J4 serves k1、k4、k6； J5 serves k3、k8、k10；
Plan 2	1.314561×10^6	5.539×10^4	4	J2 serves k2、k7、k9； J3 serves k5； J4 serves k1、k6、k8、k10；J5 serves k3、k4；
Plan 3	1.125607×10^6	6.935×10^4	4	J2 serves k2、k7、k9； J3 serves k5； J4 serves k1、k6、k8； J5 serves k3、k4、k10；
Plan 4	2.565×10^6	1.8193×10^4	5	J1 serves k6、k7； J2 serves k2、k9； J3 serves k5； J4 serves k1、k4、k8； J5 serves k3、k10；
Plan 5	2.439×10^6	1.8193×10^4	5	J1 serves k7； J2 serves k2、k8； J3 serves k5； J4 serves k1、k4、k6； J5 serves k3、k8、k10；
Plan 6	2.354×10^6	1.8194×10^4	5	J1 serves k7、k10； J2 serves k2、k9； J3 serves k5； J4 serves k1、k4、k8； J5 serves k3、k6；
Plan 7	2.323×10^6	1.8195×10^4	5	J1 serves k7； J2 serves k2、k9； J3 serves k5； J4 serves k1、k6、k8； J5 serves k3、k4、k10；
Plan 8	2.311×10^6	1.8197×10^4	5	J1 serves k7、k10； J2 serves k2、k9； J3 serves k5； J4 serves k4、k6、k8； J5 serves k1、k3；

satisfaction at any cost, choose option 4; option 4 is the distribution choice [4,2,5,4,3,1,1,4,2,5], i.e., choose the warehouse-store front warehouse alternative points J1, J2, J3, J4, J5 for the rest of the cluster point distribution.

② When company R only cares about minimizing the total cost of site selection and does not consider customer satisfaction and enterprise investment costs, option 3 is the optimal choice; option 3 is the distribution choice [4,2,5,5,3,4,2,4,2,5], i.e., choose the warehouse and store front warehouse alternative points J2, J3, J4, J5 for the rest of the cluster points to distribute goods. From the Table 11, it can be seen that the customer satisfaction penalty cost of option 3 is 69,350 is not much different from the customer satisfaction penalty cost of options 1 and 2.

③ When Company R is considering both the input cost of the enterprise and customer satisfaction, then Option 3 provides distribution services for the four warehouse-store front-loading outlets, and there is little difference in the cost of customer satisfaction from Option 4 to Option 8. However, Option 1 to 3 is the establishment of four front-loading outlets, which is a big difference from Option 4 to Option 8 in terms of the total cost. As shown in Table 11, the total cost of site selection for Option 3 is 1194957, and the total cost of site selection for Option 4, which considers the cost of customer satisfaction, is 2583193, and the total annual cost of site selection for Option 3 over Option 4 is a savings of 1388236.

VI. CONCLUSION

In this paper, we use the method of secondary clustering to cluster the crawled Gaode map POI data, the first DBSCAN clustering algorithm is used for data preprocessing to eliminate the outlier points in the original data, and the second clustering compares the K-Means algorithm with the Mean-shift algorithm, and selects the results of K-Means algorithm, which has a more uniform distribution of the cluster results, to be used as the new customer cluster points; Comparing the four classification algorithms of Naive Bayes, Support Vector Machine, Decision Tree and XGBoost, XGBoost with an AUC value of 93% is finally selected, which confirms that the algorithm has the best effect on the classification of warehouse and distribution patterns, and the classification results of the customer cluster points will be used as the alternative points for the siting of the front warehouse. The site selection model of the front warehouse establishes a multi-objective model from the perspectives of enterprise input cost and customer satisfaction, and NSGA-II is selected to solve the problem, which can solve the Pareto optimal solution set compared with the traditional weighted average, normalization, etc. According to the enterprise's own needs, it chooses the layout strategy suitable for itself and makes a scientific decision.

There are still contents that need to be further studied in this paper. Fresh food e-commerce in the warehouse location not only need to consider their own development and consumer experience, but also need to take into account the competition in the industry, in the actual layout of the competitor's cost, revenue and other conditions need to be researched, according to the actual market share of reasonable allocation, to ensure that the layout of the selected front warehouse is optimal. Regarding the research on warehouse distribution mode, due to the small volume of data, to a certain extent, it limits the effect of machine algorithms, and I hope that in the future, after the enterprises have laid out more front warehouses, there will be richer data to carry out research.

REFERENCES

[1] Z. Xiaoye, Y. Hongyue, M. Xiaoyun, and R. Guibin, "Research on the terminal node-preposition warehouse location of urban rapid logistics distribution network," *J. Shenyang Univ. Technol. Social Sci. Ed.*, vol. 13, no. 5, pp. 422–427, 2013.

[2] J. Liu, Y. Li, B. Xiao, and J. Jiao, "Coupling fuzzy multi-criteria decision-making and clustering algorithm for MSW landfill site selection (Case study: Lanzhou, China)," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 6, p. 403, Jun. 2021.

[3] F. Gocer and N. Sener, "Spherical fuzzy extension of AHP-ARAS methods integrated with modified K-means clustering for logistics hub location problem," *Expert Syst.*, vol. 39, no. 2, Feb. 2022, Art. no. e12886.

[4] W. Haili, "Improvement of K-means unsupervised clustering algorithm in big data mining," *Modern Electron. Technol.*, vol. 43, no. 19, pp. 118–121, 2023.

[5] X. Shuyang, Y. Hongfeng, P. Huazheng, and Z. Yulai, "The application of improved ant colony algorithm based on DBSCAN clustering in vehicle routing problem," *Comput. Knowl. Technol.*, vol. 16, no. 19, pp. 182–186, 2020.

[6] E. A. Neeba, S. Koteeswaran, and N. Malarvizhi, "Swarm-based clustering algorithm for efficient Web blog and data classification," *J. Supercomput.*, vol. 76, no. 6, pp. 3949–3962, Jun. 2020.

[7] B. Guanwen, L. Xiaoming, J. Yuan, S. Chunlin, D. Luxi, and T. Shaohu, "Research on taxi passenger hotspot area mining based on improved DBSCAN algorithm," *Transp. Eng.*, vol. 19, no. 4, pp. 62–69, 2019.

[8] L. Wang, P. Chen, L. Chen, and J. Mou, "Ship AIS trajectory clustering: An HDBSCAN-based approach," *J. Mar. Sci. Eng.*, vol. 9, no. 6, p. 566, May 2021.

[9] K. R. Uthayan, G. L. V. Prasad, V. Mohan, C. Bharatiraja, I. V. Pustokhina, and D. A. Pustokhin, "Clustering indoor location data for social distancing and human mobility to combat COVID-19," *Comput., Mater. Continua*, vol. 71, no. 1, pp. 907–924, 2022.

[10] M. A. Ghazanfar and A. Prügel-Bennett, "Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3261–3275, Jun. 2014.

[11] W. Weigang, C. Xiumin, J. Zhonglian, and L. Lei, "Ship trajectory classification algorithm based on weighted naive Bayes," *China Navigat.*, vol. 43, no. 4, pp. 20–26, 2020.

[12] L. Yong and Z. Weijia, "Garbage classification system based on naive Bayes classifier," *J. Liaoning Univ. Technol., Natural Sci. Ed.*, vol. 41, no. 1, pp. 49–52, 2021.

[13] X. Yingzi and R. Junling, "Based on the improved weighted complement set naive Bayesian logistics news classification," *Comput. Eng. Des.*, vol. 43, no. 1, pp. 179–185, 2022, doi: 10.16208/j.issn1000-7024.2022.01.024.

[14] W. Chen, K. Fu, J. Zuo, X. Zheng, T. Huang, and W. Ren, "Radar emitter classification for large data set based on weighted-XGBoost," *IET Radar, Sonar Navigat.*, vol. 11, no. 8, pp. 1203–1207, Aug. 2017.

[15] Z. Wu, X. Wang, and B. Jiang, "Fault diagnosis for wind turbines based on ReliefF and eXtreme gradient boosting," *Appl. Sci.*, vol. 10, no. 9, p. 3258, May 2020.

[16] Y. Liu, H. Luo, B. Zhao, X. Zhao, and Z. Han, "Short-term power load forecasting based on clustering and XGBoost method," in *Proc. IEEE 9th Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2018, pp. 536–539.

[17] P. Song and Y. Liu, "An XGBoost algorithm for predicting purchasing behaviour on e-commerce platforms," *Tehnički Vjesnik*, vol. 27, no. 5, pp. 1467–1471, 2020.

[18] L. Zhoufang, Y. Hua, and X. Zhenqiang, "The application of ant colony algorithm with clustering property in the location problem of urban vegetable logistics distribution center," *Chin. J. Agricult. Machinery Chem.*, vol. 34, no. 5, pp. 206–209, 2013.

[19] D. Yong, "Research on the location selection of fresh e-commerce O2O community stores based on customer time satisfaction," *Southwest Univ. Sci. Technol.*, 2018.

[20] S. Zhenbo, M. Wenkai, and W. Yaohua, "Optimization of pre-warehouse location based on e-commerce industry," *Sci., Technol. Eng.*, vol. 20, no. 2, pp. 681–686, 2020.

[21] P. Chron and L. Yuanxiang, "Application of improved simulated annealing algorithm in logistics distribution center location," *Statist. Decision-Making*, vol. 37, no. 9, pp. 172–176, 2021.

[22] Li Mingwei, *Research on the Location of Fresh Stores Under the Background of New Retail Based on Bi-Level Programming*. Beijing, China: Beijing Univ. Posts and Telecommunications, 2021.

[23] L. Weng, "Fresh agricultural products cold chain location selection in context of big data," *J. Physics: Conf. Ser.*, vol. 1631, no. 1, Sep. 2020, Art. no. 012122.

[24] B. Shu, F. Pei, K. Zheng, and M. Yu, "LIRP optimization of cold chain logistics in satellite warehouse mode of supermarket chains," *J. Intell. Fuzzy Syst.*, vol. 41, no. 4, pp. 4825–4839, Nov. 2021.

[25] L. Xiangtao, *A Company Beijing Pick-Up Network Layout Research*. Beijing, China: Beijing Jiaotong Univ., 2015.

WEI XU was born in 1979. He received the bachelor's degree in management and the D.Eng. degree from the Shandong University of Science and Technology, in 2001 and 2015, respectively. He is currently with the Shandong University of Science and Technology, where he is currently a Professor with the Department of Transportation Engineering, School of Transportation. His research interests include logistics system planning and design, traffic data analysis, logistics, and transportation.

CHAO WANG was born in 2000. He received the bachelor's degree in engineering from Xinjiang Agricultural University and the master's degree from the Shandong University of Science and Technology, where he is currently pursuing the Graduate degree in transportation. His research interests include logistics system planning and design, big data analysis, logistics, and transportation.

LEI XING was born in 1989. He received the M.S. and Ph.D. degrees in engineering from Dalian Maritime University, in 2016 and 2021, respectively. He is currently with the Shandong University of Science and Technology, where he is currently a Lecturer with the Department of Transportation, School of Transportation. His research interests include logistics system planning and design, supply chain management, logistics, and transportation.

NAN LI was born in 1998. He received the bachelor's degree in engineering from the Nanjing University of Science and Technology, in 2020, and the master's degree from the Shandong University of Science and Technology, in 2023. He is currently with the Qingdao Branch of Shandong Road and Bridge. His research interests include logistics system planning and design, big data analysis, and supply chain management.