scientific reports

Check for updates

OPEN

Optimization of frozen goods distribution logistics network based on k-means algorithm and priority classification

Jianli Shi

Maintaining the quality and integrity of frozen goods throughout the supply chain necessitates a robust and efficient cold chain logistics network. This research proposes a machine learning-based method for optimizing such networks, resulting in significant cost reduction and resource utilization improvement. The method employs a three-phase approach. First, K-means clustering groups sellers based on their geographical proximity, simplifying the problem and enabling more accurate demand prediction. During the second phase of the proposed method, Gaussian Process Regression models predict future sales volume for each seller cluster, leveraging historical sales data. Finally, the Capuchin Search Algorithm simultaneously optimizes distributor location and resource allocation for each cluster, minimizing both transportation and holding costs. This multi-objective approach achieved a 34.76% reduction in costs and a 15.6% reduction in resource wastage compared to the existing system. This novel method offers a valuable tool for frozen goods distribution networks, with advantages such as considering multiple goals for optimization, focusing on demand prediction, potential for reduced complexity, and focusing on managerial insights over compared methods.

Keywords Cold chain logistics, K-means clustering, Gaussian process regression, Capuchin Search Algorithm

In order to achieve budgetary requirements while minimizing expenses and preventing quality loss during storage and distribution, cold chain logistics is crucial in maintaining the appropriate commodities at the right time. The cold chain is advised to maintain the proper temperature for perishable goods during the distribution process as a crucial component of the logistic system[1–3]. The perishable goods must be kept in a freezer box equipment due to the storage conditions. Frozen food is one form of perishable product with a short shelf life and limited time of sale. It is distributed by a freezer box truck, which has a greater operating cost and uses more fossil fuel than a standard vehicle. Therefore, it is thought that the effectiveness of the distribution of frozen goods has a substantial impact on both operating expenses and retail sales. The drivers must also adhere to the consumers' and stores' time-window requirements in order to boost service satisfaction[4]. A distributor that sells items with a limited shelf life, like milk, ice cream, and lunch boxes, levies a late delivery fee. Customers have higher expectations for the quality of fresh products as living standards rise[5].

The quality of fresh products is significantly influenced by the temperature; specifically, high temperatures hasten product degradation[6]. Due to its capabilities of maintaining product quality via low temperature, cold chain logistics has emerged as the primary way of distributing fresh goods in this setting[7]. The quality of fresh food will decline as time goes on in addition to temperature[8]. Because cold chain logistics has higher standards than traditional logistics, cold chain logistics businesses face enormous hurdles in the area of customer service. Nowadays, businesses no longer have a clear price edge, so they search for new competitive advantages to boost customer satisfaction. Only by doing this will they be able to stand out from the competition in the market for cold chain logistics[9]. Therefore, cold chain logistics companies should thoroughly investigate ways to improve logistics services to raise client satisfaction[10].

The key to evaluating the current service quality of cold chain logistics businesses is customer satisfaction measurement, which can improve communication between customers and businesses[11]. Additionally, measuring allows for the identification of the critical variables influencing customer satisfaction, which helps to reveal the advantages and disadvantages of the cold chain logistics firm and enhance logistical operations[12].

School of Management Science and engineering, Chongqing Technology and Business University, Chongqing, Chongqing 400061, China. email: shjl20043528@163.com

The act of distributing products or goods from a manufacturer to a customer is known as distribution. The distributor transports the goods using a vehicle during the technical execution of the procedure. A sound plan is necessary to reduce distribution costs. Choosing the paths taken by the vehicles is part of this strategy. The vehicle routing problem (VRP) is the name given to the issue. The VRP has been addressed using a variety of techniques and is regarded as a combinatorial optimization problem[13,14].

A few of the distributor's trucks will be used in the technical execution of the product distribution to transport the goods to clients. Distributors must choose the best route for the distribution of goods given the various consumer bases. Minimizing the cost of distribution is the major goal of the solution. By reducing distribution routes and the utilization of used vehicles in the distribution process, a minimum distribution cost may be attained. Along with these, it's important to take the vehicles' capacity into account[15]. The benefits will be gained to a greater extent the higher the quality of the distribution procedure.

For frozen goods delivery, cold chain logistics network optimization is essential to preserving product integrity and quality across the supply chain. But conventional approaches to distributor location and resource allocation are frequently too simple to address intricate aspects including sellers' geographical dispersion, shifting demand patterns, and the requirement to keep holding costs as low as possible. In order to tackle this problem, the research suggests a new, three-phase machine learning and optimization-based strategy. This strategy seeks to significantly lower costs and enhance the use of resources in frozen food cold chain logistics networks.

The motivation of the current research is to answer the following questions:

1. How can k-means clustering be effectively utilized to group geographically proximate sellers in a cold chain logistics network to improve demand prediction accuracy?
2. Can Gaussian Process Regression models leverage historical sales data to predict future sales volume for seller clusters, enabling better resource allocation within the network?
3. Does a combined approach using k-means clustering, Gaussian Process Regression, and the Capuchin Search Algorithm lead to a more efficient distribution network design compared to traditional methods, considering both transportation and holding costs?

This research merges machine learning and optimization techniques to handle complexities like seller location, demand variation, and cost minimization. The innovation lies in the combined use of k-means clustering for location-based grouping, Gaussian Process Regression for demand prediction, and the Capuchin Search Algorithm for simultaneous distributor placement and resource allocation, minimizing both transportation and holding costs. This multi-objective approach ushers in a new era of cold chain logistics optimization, with the potential for significant cost reductions and improved resource utilization. In decision, the list of our most significant contributions is as follows:

- Establishing seller clusters based on geographic information.
- Projecting the number of sales that each cluster's merchants will make.
- Making use of resource allocation and distributor placement for each cluster.

The paper unfolds as follows: In Sect. "Research background", we review the related works. The proposed method in Sect. "Research methodology", followed by Experimental results in Sect. "Experimental results" and in Conclusion in Sect. "Conclusion".

Research background

The current section aims at presenting a literature review of the recent developments in supply chain networks with an emphasis on those that are suitable for the distribution of frozen products. Some of the areas that are often discussed in relation to the subject include vehicle routing, production distribution planning, demand forecasting, big data and sustainability. Hence, the review seeks to find out the gaps in the current literature and lay down a firm research framework for subsequent research on the efficient distribution of frozen goods through logistics networks.

A. Vehicle Routing and Distribution Network Design.

In this part, the issues and progress of vehicle routing and distribution network design are discussed in reference to supply chain management. Zhang et al.[16] explored the use of genetic algorithms in solving vehicle requirements planning problems, focusing on the optimization of complex logistics distribution paths and the study of temporary customer orders. Gámez-Albán[17] suggested a mixed-integer multi-period programming paradigm to reduce logistics network costs for a Colombian veterinary products distributor. The logistics network model's layout takes into account the cost of out-of-stock inventories as well as the opening and closure of facilities over the planning horizon specified by the company. Rodríguez et al.[18] developed a distribution logistics model based on a case study to ensure the lowest logistics cost. It considered demand, supply, factory capacity, distribution center sites, and stores. The model identified customer restrictions and simulated their behavior using data processing. The papers discussed in this section clearly indicate that the challenges posed by the vehicle routing problems are becoming more difficult and require new and efficient solutions to satisfy customers' requirements and tackle practical constraints. future research could consider developing machine learning and artificial intelligence algorithms that would allow for more efficient vehicle routing decisions in frozen food distribution networks with regard to demand fluctuations, product priority, and resource availability.

B. Production-Distribution Planning and Demand Forecasting.

This subsection focuses on the importance of production-distribution planning and demand forecasting in the supply chain performance. Ariafar et al.[19] created a mathematical model for a production-distribution dilemma that uses fuzzy set theory to allocate resources and select the best contractors for product delivery.

Utilizing a three-step solution technique, the model was evaluated in a mineral water bottling facility. Real-world scenarios were used to show the model's viability and validity, and even with different -cut levels, the decision-maker's pleasure remained constant. This strategy is essential for the effective design and operation of supply chains in the modern, cutthroat corporate climate. Li[20] proposed that, particularly in hub and width type networks, the use of predictive data transmission technology (PDTT) in cold chain logistics may greatly retain the freshness of fruits and vegetables, cut transportation costs, and lengthen the shelf life of food.

Leung et al.[21] presented a machine learning predictive model for predicting near-realtime order arrival in e-commerce distribution centers. It gets around the drawbacks of longer time horizons for operations management. The approach makes use of moving average, volatility, and autoregressive components as well as historical order arrival data from downstream retailers. For better supply chain management decision-making, future research can combine optimization or heuristics approaches with predictive methodology.

Cai et al.[22] proposed a spatial feature fusion and grouping strategy for e-commerce commodity demand forecasting. It creates a neural network prediction model that demonstrates the beneficial effects of multimodal data, customer testimonials, and consumer profiles on demand forecasting. Ablation experiments are used to show the superiority of spatial feature fusion, and the e-commerce product dataset is used to test the model's efficiency and superiority.

The research studied in this subsection shows the need to incorporate efficient methods of forecasting and optimization to aid in the effective use of the resources and delivery of the products on time. future research could be directed to the enhancement of demand forecasting models for frozen products that takes into consideration factors like temperature changes, seasonal changes, and occasions.

C. Big Data, Cloud Computing, and Customer Satisfaction Analysis.

This section aims at exploring the effects of big data and cloud computing in supply chain management especially the satisfaction analysis of the customers. Chen et al.[23] concentrated on the utilization of big data and cloud computing in cold chain logistics. Using real-time traffic data, it examines the cost and duration of refrigerated truck distribution. To solve the optimization model, the study makes use of cloud computing's parallel programming mode. The strategy is effective, with shorter execution times for additional processors, according to experimental results.

Lim et al.[24] recommended a latent dirichlet al.location (LDA) model to identify eight customer-related issues, calculate sentiment scores in user-generated reviews, and conduct regression analysis to discover key factors influencing customer sentiment. Results demonstrate that factors such as timeliness, cost, cold chain transportation, quality, error handling, and customer service personnel have a big impact on feeling good.

In this subsection, the reviewed studies highlight the benefits of big data and cloud computing, such as increasing supply chain transparency, improving decision making and increasing customer satisfaction. the future research may focus on how to employ more sophisticated analytical tools to analyze the big data relevant to frozen goods distribution including delivery rate and customer satisfaction data.

D. Delivery Time Uncertainty and Network Efficiency.

This section focuses on the issues arising from uncertain delivery time and examines the best ways of minimizing network problems. Wang et al.[25] developed a bi-objective programming model to manage delivery time uncertainty in logistics operations, optimizing overall operating cost and delivery vehicle number, and thereby increasing urban logistics and intelligent transportation network efficiency. The delivery time uncertainty is an essential factor in supply chain and that optimization methods are needed to minimize the effects of this uncertainty. In future research, more efforts could be made to build models that are more precise in estimating the delivery time risk and uncertainty of frozen products with reference to some factors such as traffic jam, weather condition, and equipment breakdown.

E. Resilient and Sustainable Supply Chain Design.

This section is dedicated to the emerging role of resilience and sustainability in the supply chain network design. The limitations of conventional k-means clustering in distribution center location problems are examined by Lin et al.[26]. To get over obstacles and improve distribution center placement, they suggest a brand-new Fruit-fly Optimization K-means (FOA K-means) algorithm. The simulation-optimization techniques for building robust supply chain networks in the face of disturbances are reviewed by Tordecilla et al.[27]. They classify previous studies and find ways to combine different criteria and hybrid approaches to deal with ambiguities.

A multi-objective fuzzy robust optimization technique is presented by Nayeri et al.[28] for creating closed-loop supply chain networks that are sustainable. Their approach acknowledges the inherent uncertainties and takes into account the financial, environmental, and social repercussions. In a case study on the manufacturing of medical devices, Hasani et al.[29] suggest a multi-objective optimization model for creating a resilient and environmentally friendly global supply chain network. Their approach uses a novel hybrid heuristic to find an efficient solution while balancing resilience issues with economic and environmental goals.

By building supply chain network optimization models that specifically consider labor availability and capacity constraints as crucial variables, Nagurney[30] offers a distinctive viewpoint. This study emphasizes how crucial it is to take people into account while building resilient networks, especially in light of disruptions like the COVID-19 pandemic. In general, supply chain network optimization is a field that is always changing as a result of researchers creating cutting-edge methods to tackle challenging problems and increase sustainability, resilience, and efficiency.

The papers considered in this subsection show that there is a need for creating new strategies for constructing robust and environmentally friendly supply chain networks that are capable of managing disruptions. as for the future research, it is possible to consider the further construction of the quantitative indicators and models for evaluating the cold chain logistics network sustainability and robustness based on the product perishing rate, temperature sensitivity, and energy consumption. Table 1 summarizes the literature review.

Ref.	Year	Objective	Method	Limitation
Zhang et al.[16]	2022	Optimize distribution paths in cold chain logistics	Genetic Algorithm	Limited to solving complex path problems, may not be suitable for all network designs
Gámez-Albán et al.[17]	2017	Reduce logistics network costs	Mixed-integer multi-period programming	Relies on accurate estimations of future costs and inventory levels
Rodríguez et al.[18]	2022	Minimize logistics costs in distribution network	Distribution logistics model with data processing	Relies on the accuracy and completeness of the processed data
Ariafar et al.[19]	2014	Allocate resources and select contractors for production-distribution	Fuzzy-based mathematical model	Requires expertise in fuzzy set theory and may not be computationally efficient for large-scale problems
Li [20]	2021	Improve product freshness, reduce transportation costs, and extend shelf life in cold chain logistics	Predictive data transmission technology (PDTT)	Relies on reliable and extensive data transmission infrastructure
Leung et al.[21]	2020	Predict near-real-time order arrival in e-commerce distribution centers	Machine learning model	Reliant on the quality and historical accuracy of training data
Cai et al.[22]	2021	Improve e-commerce commodity demand forecasting	Spatial feature fusion and grouping strategy with multimodal data	Requires access to diverse and high-quality customer data
Chen et al.[23]	2020	Optimize refrigerated truck distribution costs and durations in cold chain logistics	Big data and cloud computing with real-time traffic data	Requires significant investment in big data infrastructure and expertise
Lim et al.[24]	2021	Identify customer concerns in cold chain logistics and explore factors influencing customer satisfaction	Latent Dirichlet allocation (LDA) for sentiment analysis of user reviews	Relies on the representativeness and volume of user reviews available
Wang et al.[25]	2020	Manage delivery time uncertainty and improve efficiency in urban logistics and intelligent transportation networks	Bi-objective programming model	Relies on accurate estimations of future demand and traffic conditions
Lin et al. [26]	2022	Improve distribution center location for better network efficiency	Fruit-fly Optimization K-means (FOA K-means) algorithm	Limited to clustering problems and may require adaptation for other network design aspects
Tordecilla et al. [27]	2021	Design resilient supply chain networks under disruptions	Simulation-optimization methods	Requires significant computational resources and expertise for model development
Nayeri et al.[28]	2020	Design sustainable closed-loop supply chain networks	Multi-objective fuzzy robust optimization	Requires expertise in fuzzy set theory and may be computationally expensive for complex networks
Hasani et al.[29]	2021	Design a green and resilient global supply chain network	Multi-objective optimization model with a novel hybrid heuristic	Relies on the accuracy of data used to define environmental and social objectives
Nagurney[30]	2021	Design resilient supply chain networks by considering labor availability and capacity limitations	Supply chain network optimization models with labor variables	Requires access to detailed labor data and may require model adaptations for specific industries

Table 1. Summary of the literature review.

Feature	Description	Count/Format
Sellers	Unique sellers in the dataset	910
Sales Records	Total number of bulk sales records	43,420
Time Period	Period covered by the data	October 2021 - October 2022 (1 year)
Products	The number of products being sold in cold chain food distribution	24
Vehicles	Number of assignable vehicles in the distribution network	82

Table 2. Dataset specifications.

Research methodology

This section outlines the proposed strategy for optimizing the logistics distribution network for frozen products using a combination of machine learning and optimization techniques. The proposed strategy leverages unsupervised and supervised machine learning techniques to decompose the problem and provide an approximate optimal solution, which is achieved by analyzing sales data within the distribution network. Therefore, in the following sections, we will first describe the assumed system model and the data structure used in this research, and then present the proposed methodology.

Dataset

The research utilizes a dataset consisting of bulk sales records of frozen products in the Guangdong province over a one-year period. This dataset includes 43,420 records related to bulk sales made by 910 sellers from October 2021 to October 2022. Each data record contains the seller's identifier, seller's location information, sales date and time information, and sales volume. The seller's identifier is a unique natural number assigned to each seller. Location information of the seller is described in terms of longitude and latitude coordinates, specifying the coordinates of the seller's location for receiving goods from the distributor. Sales time information includes the date (in day-month-year format) and the sale time (in a 24-hour format). Finally, the sales volume feature in each data record represents the frequency of the product sold in that transaction. Table 2 lists the dataset specifications.

Given the current research objective, features such as the seller's identifier and product sale time are not considered, and each data record is represented as a four-element set $\{Longitude, Latitude, Date, Amount\}$.

Proposed model

In this research, a logistics distribution network for frozen products is considered, consisting of six main components: 1- Producer, 2- Product Transport, 3- Distributor, 4- Store/Seller, 5- Product, and 6- Buyer. The product and buyer components are implicitly modeled by other components. The assumed system model is depicted in Fig. 1. In this model, the product producer sends the products using specialized transportation to distributors. Each transport vehicle has limited capacity and incurs a cost on the distribution network proportional to the distance traveled. Each distributor also has a specific location and limited capacity, and its operations come with a fixed cost proportional to its capacity and location. The role of distributors in the logistics distribution network is to maintain products under suitable conditions and distribute them among sellers based on generated demands. Each distributor utilizes the assigned transportation vehicles to transport and deliver products to sellers. Additionally, the coverage area of each distributor is determined through receiving stores for that distributor, and it is assumed that the coverage areas of distributors in the assumed system model do not overlap. Thus, the resources required for each seller are provided only by one distributor. In the assumed system model, each vehicle transports products related to only one distributor. Finally, sellers act as intermediaries between buyers and the distribution network, and they can use various sales platforms (online/in-person) for transactions with buyers. Each seller has a specific location for receiving products from the distributor.

In this research, the optimization problem of a logistics distribution network for frozen products is tackled through determining the distributor's location and allocating product transportation resources. In contrast to conventional logistics optimization issues, this work exhibits some distinctive features:

- Joint Cost Minimization: The model optimizes more than just distributor holding costs or transportation expenses. It uses a multi-objective strategy to decrease distributor holding costs (based on capacity and location) and transportation costs (proportionate to journey distance) at the same time.
- Demand Prediction Integration: This study takes into account fluctuations in demand across geographically distributed suppliers, in contrast to many traditional logistics optimization problems. The model groups vendors according to geography using k-means clustering (explained in Sect. "Clustering sellers based on Location Information"), which enables more precise demand forecast using Gaussian Process Regression (Sect. "Predicting the sales volume of sellers in each cluster"). This makes it possible to allocate resources to various seller clusters in a more sophisticated manner.

The designated seller locations define each distributor's coverage region. The model makes sure that the coverage zones of the distributors do not overlap. By ensuring that every seller obtains resources from a single, approved distributor, this streamlines resource allocation and lowers the possibility of delivery conflicts.

This research adds to the optimization of cold chain logistics networks by taking demand changes and cost minimization into account, all while addressing these unique characteristics. For the distribution of frozen items, this multi-objective strategy with integrated demand prediction provides a more workable and effective solution.

In the continuation of this section, the details of the proposed method for optimizing the distribution network of frozen products using machine learning techniques and optimization are explained. The goal of this approach is to find the optimal location of distributors in the logistics distribution network and allocate resources to them

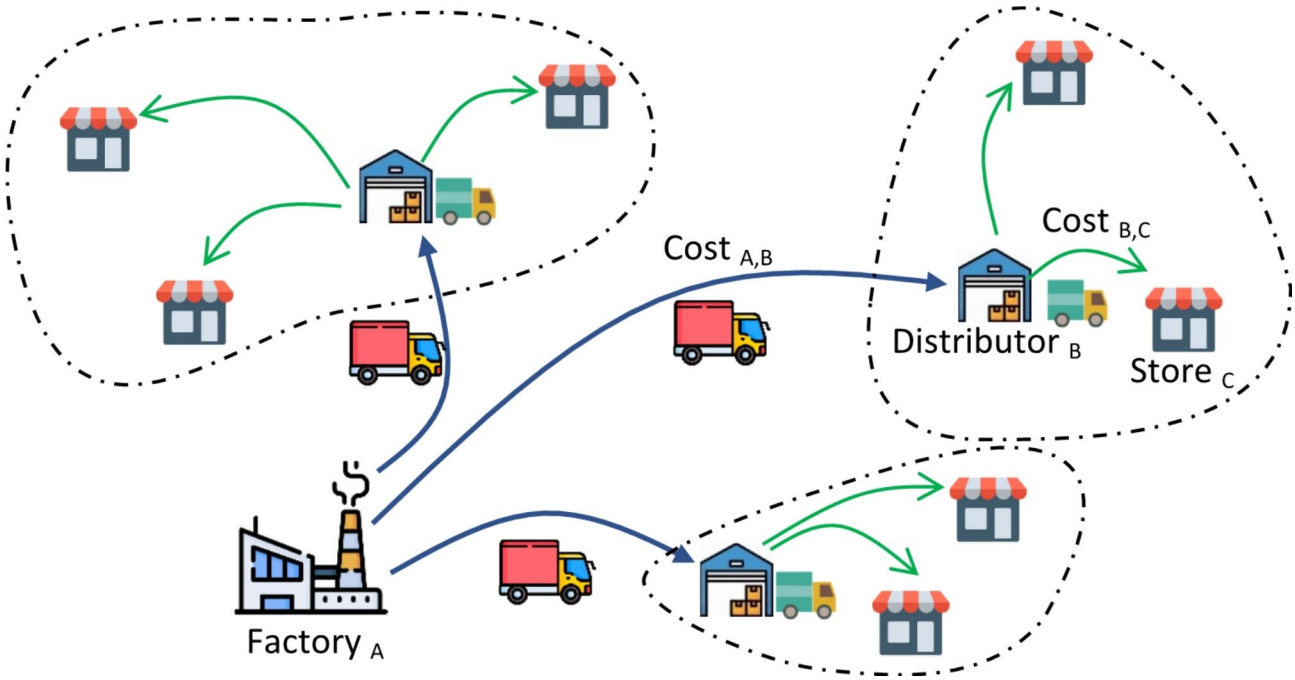

Fig. 1. Proposed system model.

in such a way that minimizes the costs imposed on the distribution network. The proposed model to achieve this goal consists of three main phases:

1. Clustering of sellers based on location information.
2. Prediction of sales volume for sellers in each cluster.
3. Distributor location and resource allocation for each cluster.

The diagram of the stages of the proposed method is illustrated in Fig. 2.

In the first phase of the proposed method, the existing sellers are divided into categories based on their location information. This process is carried out using the K-means clustering algorithm. This aims to not only reduce the complexity of the problem by dividing it but also to obtain more accurate models for predicting product demand in the distribution network. The result of the first phase will be a set of clusters, each of which contains one or more sellers.

In the second phase, a Gaussian Process Regression (GPR) model is utilized to predict future sales volumes for the sellers within each cluster. To achieve this, the sales data related to sellers in each cluster are organized into time series data, and this data is used to train the GPR model. Subsequently, the GPR models are used to predict the sales volume by sellers located in each cluster. The information obtained from clustering sellers and predicting their sales volume serves as input for the third phase of the proposed method.

In this phase, the Capuchin Search algorithm is employed to determine the optimal location for distributors corresponding to each cluster and allocate resources to each one. This optimization model aims to address the problem by simultaneously considering both transportation cost and distributor holding cost.

Clustering sellers based on Location Information

The proposed method begins with clustering the existing sellers based on their geographical location information using k-means algorithm. Utilizing k-means clustering strategy for the sellers in the proposed method can have two significant advantages of reduced complexity and enhanced accuracy.

Firstly, forming clusters of nearby sellers reduces the complexity of the research problem. This is because the resource allocation problem in the distribution network is cost-based, and the distribution distances play a fundamental role in determining the total cost. In this case, assigning a common distributor to neighboring sellers is straightforward. Therefore, through clustering sellers, the problem of "determining a distributor for

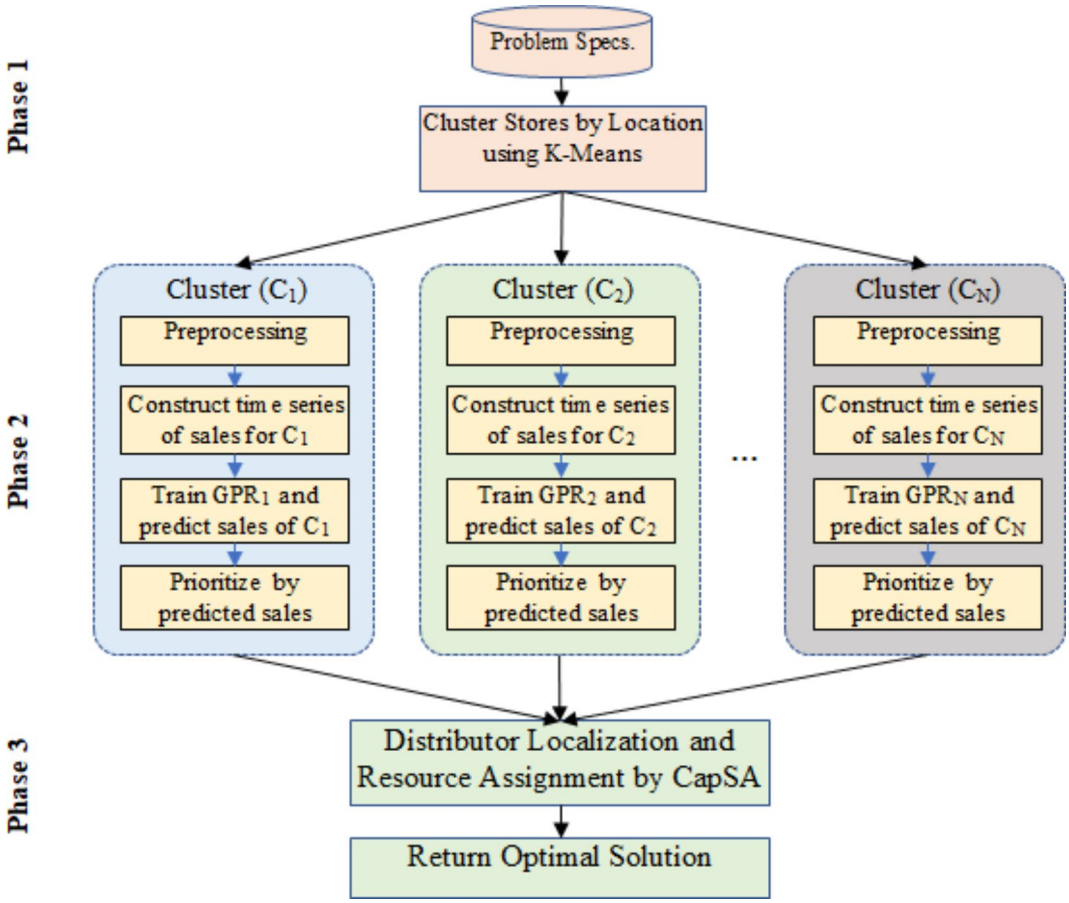

Fig. 2. Diagram of the Proposed Method Stages.

each seller" can be transformed into a set of problems of "determining a distributor for a set of neighboring sellers."

Secondly, the sales pattern of sellers can be related to their location-based information (due to characteristics such as population, buying patterns, events, etc.). For more accurate sales pattern modeling, sellers can be categorized based on geographical information, and then a separate learning model can be used to model the sales pattern for each category. These advantages have led to the use of the K-Means algorithm in the first phase of the proposed method.

The clustering of sellers is solely based on their geographical location information. As per the data structure described in Sects. "Research methodology" – "Introduction", the seller's location is described using two attributes, Longitude and Latitude. By extracting these two features for each seller, a data record is formed. The K-Means clustering algorithm is a fundamental method for many other clustering techniques. This algorithm follows an iterative process and repeats the following steps for a fixed number of target clusters:

1. Determining the centroids (center points) of each cluster, which is determined by calculating the average of the members assigned to that cluster.
2. Adjusting the members of each cluster based on the minimum distance of each data point to the cluster centroids.

This iterative process continues until convergence, and the result is a set of clusters where each seller belongs to a specific cluster based on their geographical proximity.

In the proposed method, N sellers provide input to the K-means algorithm based on their geographical coordinates of longitude and latitude. Initially, a random center is assigned to each cluster. Then, the distance of each seller from the cluster centers is calculated, and each seller is assigned to the cluster with the minimum distance to its center. By changing the clustering structure, new cluster centers are recalculated (as the average of the points assigned to that cluster), and the process of adjusting the cluster structure is repeated. This process continues until the clustering structure obtained from two consecutive iterations remains relatively unchanged. It is worth mentioning that, based on the description of geographical coordinates using longitude and latitude, the proposed method utilizes the Haversine distance metric to evaluate the distance between two points. A sample clustering result for some of the stores in the database is shown in Fig. 3. In this process, stores are grouped

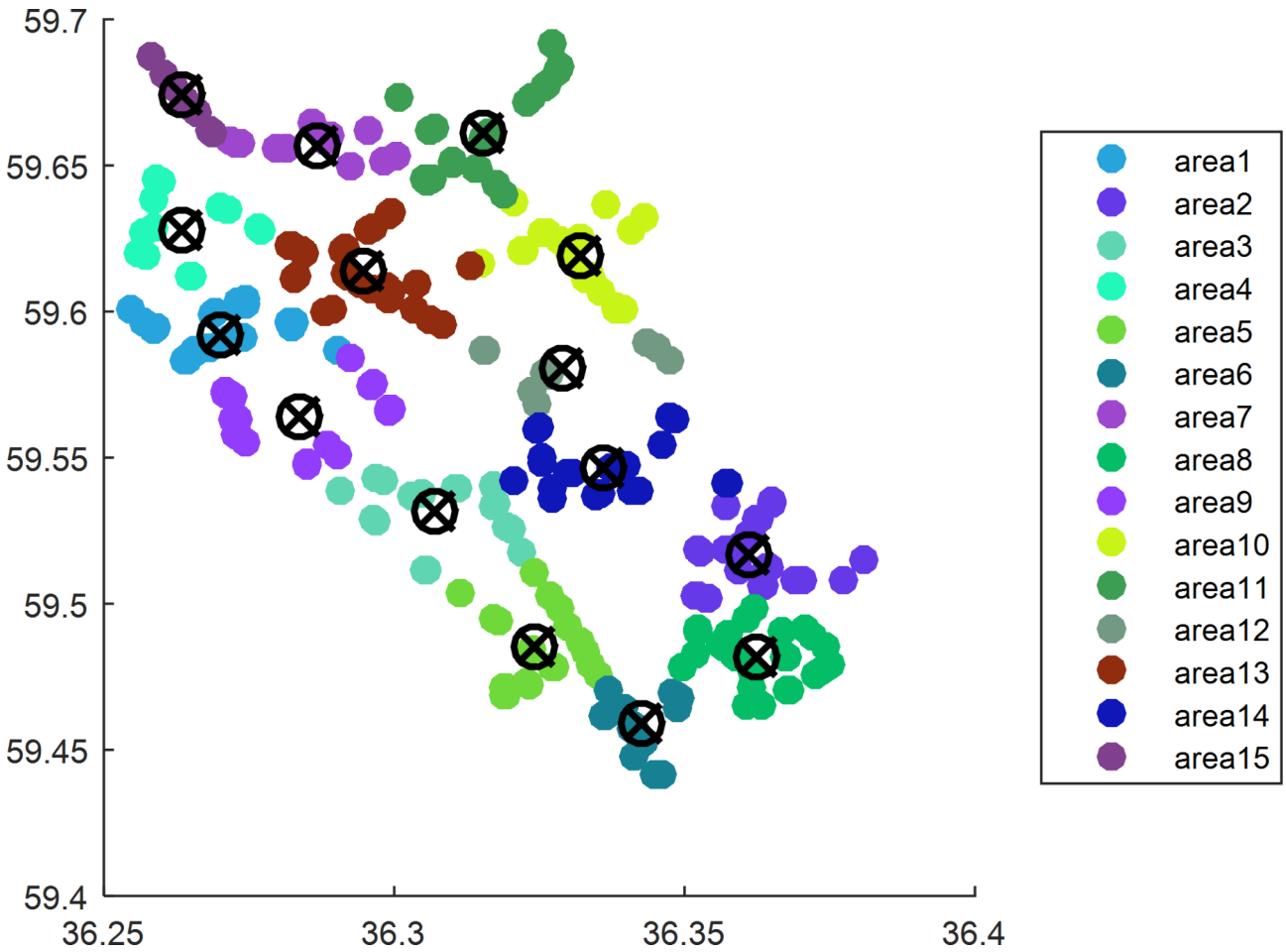

Fig. 3. An example of store clustering results based on location using the K-Means algorithm.

into clusters based on their geographical proximity, and the centers of these clusters are iteratively updated until convergence. This clustering step helps reduce the complexity of the distribution network optimization problem and provides a basis for further analysis and resource allocation in the proposed logistics distribution optimization method.

In Fig. 3, the members of each cluster are represented as points of the same color, and the centers of each cluster are represented as "U" shapes. The results of clustering stores serve as input for the second phase of the proposed method. Furthermore, these clustering results and their centers provide the necessary data for the optimal location allocation of distributors in the third phase of the proposed method.

Predicting the sales volume of sellers in each cluster

In the proposed method, after clustering the stores, a time series is created to describe the weekly sales volume of each cluster. The first step in the second phase of the proposed method is to transform the weekly sales information of each cluster into a format that can be processed by machine learning models. To preprocess the data of each cluster, records containing missing values are first disregarded, and then sales date information is organized into weeks elapsed from the beginning of the year. Subsequently, the sales volume values for each week in a cluster, such as cluster i, are calculated by summing the sales volume values of the sellers belonging to cluster i. This process allows the sales volume values in each cluster to be described as {x(1), x(2), …, x(N)}, where x(j) represents the total sales volume of sellers in cluster i for week j. The input sequence for the system to understand the sales volume of a cluster in the k-th week can be represented as follows:

$$X\left(k\right)=\left[x\left(k-n\times\ t\right),\ldots\ ,x\left(k-t\right),x(k)\right] \tag{1}$$

In the equation above, 'n' represents the size of the sequence, and 't' is the time difference between each sequence element. For example, considering n = 3 and t = 5, the sequence derived from Eq. 1 for time 'k' would be X(k)=[x(k-15), x(k-10), x(k-5), x(k)]. This means that for modeling a sales volume prediction system, the input sequence not only includes the sales volume in the current week but also contains the sales volumes in regular past time intervals (based on weeks). The goal of a sales volume prediction model is to provide an appropriate estimate of the sales volume in week k + 1 given the input x(k).

$$y\left(k+1\right)=f\left(x\left(k\right)\right) \tag{2}$$

In the equation above, 'y' represents the predicted sales volume. The goal of this research is to provide a suitable approximation for the function 'f' using the GPR model. The following explanation outlines the computational model of GPR for predicting the sales volume of each cluster and modeling 'f'. Two main advantages made GPR a suitable machine learning algorithm for this purpose:

- Taking care of Demand Data Non-linearities: Non-linear connections between variables like as geography, seasonality, and product type are frequently observed in real-world sales data. When modeling non-linear relationships in the historical sales data of each seller cluster found by k-means, GPR is an effective tool. Compared to linear models, this enables a more nuanced understanding of demand changes, which could result in better decisions about the allocation of resources.
- Quantification of Uncertainty: One of GPR's key advantages is its capacity to measure the degree of uncertainty around its demand projections. Given the inherent fluctuation in demand, this information is essential because it enables us to evaluate the degree of confidence in our forecasts and make better educated judgments regarding the allocation of resources.

Unlike many widely used supervised machine learning algorithms that learn precise values for each parameter in a function, the Bayesian approach infers probability distributions over all possible parameter values. We assume a linear function as y = wx + e. The Bayesian approach works by specifying a prior distribution over the parameter 'w' and updating the probabilities based on evidence (i.e., observed data) using Bayesian rule. It is described as follows[31]:

$$P\left(w|y,X\right)=\frac{P\left(y|X,w\right)p\left(w\right)}{p\left(y|\ X\right)}\ \rightarrow\ posterior=likelihood\times\ \frac{prior}{marginal\ likelihood} \tag{3}$$

In the equation provided, P(w | y, X) represents the probability that the value of 'w' can provide a good estimate of the next 'y' based on observations 'X'. This probability is represented as the posterior and will be further referred to as future probabilities. Additionally, P(y | X, w) represents the probability of 'w' matching previous observations 'X' and 'y' (training data). This probability is determined as the likelihood. Furthermore, 'p(w)' represents the probability of 'w' occurring in previous observations or, in other words, the training data. This probability will be referred to as prior probability in the following text.

Gaussian Process Regression (GPR) is a non-parametric model, which means it is not limited to a specific functional form. Instead of calculating the probability distribution of specific parameters for a particular function, GPR calculates probability distributions over all acceptable functions based on the data. This involves specifying a prior probability distribution in the function space and updating probabilities using observed data to compute future probabilities. Consequently, we can predict the distribution at desired points[31].

In the GPR process, a prior Gaussian process is considered, which can be determined using a mean function and covariance function. Specifically, a Gaussian process is like an infinite-dimensional Gaussian distribution where sets of data labels share a joint Gaussian distribution. In this prior Gaussian process, prior knowledge

can be incorporated into the function space by selecting mean and covariance functions. To calculate future probability distributions (the target variable), data and test observations are conditioned on. Since a Gaussian process is chosen as the prior probability distribution, the distribution of the predicted variable can be computed, resulting in a Gaussian distribution described by mean and covariance[31].

As previously mentioned, the proposed method uses a separate GPR model to model the sales patterns of products in each cluster, and this model is trained solely based on the weekly sales records by the vendors within that cluster. After training each GPR model, it is used to predict the sales volume for the cluster in future time intervals. The predicted values for different clusters are sorted, and based on the predicted sales volume, each cluster is assigned a priority value. The information obtained in this phase is used as input in the third phase of the proposed method.

Location of distributors and resource allocation for each cluster

In the third phase of the proposed method, the location of distributors and the allocation of resources to them are determined using the optimization strategy called Capuchin Search Algorithm (CapSA).

The choice of the CapSA for distributor placement and resource allocation within our proposed method stems from its inherent strengths in handling NP-Hard optimization problems. Our research objective demands simultaneous optimization of two competing objectives: minimizing transportation cost and minimizing distributor holding cost. While traditional optimization techniques often struggle to balance multiple objectives effectively, CapSA offers distinct advantages in this regard:

- Balancing Exploration and Exploitation: CapSA maintains a healthy balance between exploring the search space for potentially superior solutions and exploiting promising regions identified during the search process. This balance prevents stagnation in local optima and allows for global search, improving the likelihood of finding an optimal solution that satisfies both cost and resource utilization objectives.
- Data-Driven Parameter Tuning: Instead of relying on pre-defined parameters, CapSA can dynamically adjust its search behavior based on the specific problem characteristics. This data-driven approach ensures the algorithm adapts to the nuances of our distribution network optimization problem, leading to more efficient and effective search.
- Computational Efficiency: Compared to other optimization algorithms, CapSA exhibits favorable computational efficiency, particularly for problems with a moderate number of decision variables. This efficiency is crucial for our application, allowing for practical implementation within time and resource constraints.

For each cluster obtained from the first phase of the proposed method, one or more candidate locations for deploying the distributor(s) are identified. Additionally, optimal resource allocation, including vehicles for transporting products, is carried out for each deployed distributor in this phase. These processes are performed in the proposed method with the goal of minimizing the overall logistics distribution network cost.

To elaborate further, let's break down the problem and encoding of solution vectors. Then, after formulating the optimization problem, we'll discuss the steps involved in CapSA for solving this problem.

Consider a frozen product distribution network problem involving s vendors organized into K clusters. The number of available delivery vehicles, denoted as C, is assumed for allocation to distribution centers in this problem. The objective of this phase of the proposed method is to allocate $x \geq 1$ distributors to each cluster of vendors and allocate $c \geq 1$ delivery vehicles to each deployed distributor. The deployment location of distributors is determined based on a set of candidate locations for each cluster. In other words, the location of each distributor is not freely determined, and its position for cluster i is specified based on one of the pre-defined positions p_i for that cluster. Consequently, the optimization parameters in the current problem will include:

1. Determining the deployment or non-deployment of a distributor in $P = \sum_{i=1}^{K} p_i$ locations for the K clusters, which is represented as a binary string of length P.
2. Determining how to allocate C delivery vehicles to the established distribution centers, represented as a natural number vector of length C.

Thus, the total number of optimization variables in the proposed algorithm is P + C. Accordingly, each solution vector will have a length of P + C. In each solution vector, the first P elements correspond to the candidate locations for deploying distribution centers in vendor clusters and are represented as a binary string. In this binary string, a value of 1 indicates the deployment of a distributor at the corresponding candidate location, while 0 represents the non-deployment of a distributor at that candidate location. The second part of each solution vector is represented as a natural number vector of length C. Each element in this part of the solution vector corresponds to one of the delivery vehicles, and the number in each element indicates the distributor's identifier to which the vehicle is allocated. Furthermore, if an element corresponding to a vehicle has a value of 0, it means that the vehicle is not assigned to any distributor.

For example, Consider a problem with 7 candidate locations ($P = 7$) for 3 clusters ($K = 3$) and 4 vehicles ($C = 4$). In this example, first three candidate locations $\{p_1, p_2, p_3\}$ are located in coverage region of cluster 1, Also $\{p_4, p_5\}$ and $\{p_6, p_7\}$ are located in coverage region of clusters 2 and 3, respectively. In this case, each solution vector is coded by 7 binary variables in addition to 4 numeric variables. Solution vector of {*1*, *1*, *0*, 1, 0, *0*, *1*, *1*, *3*, *1*, *2*} represents:

- Cluster 1: Deploys two distributors at locations p_1 and p_2.
- Cluster 2: Deploys a distributor at location p_4.
- Cluster 3: Deploys a distributor at location p_7.

- Vehicle allocation: Vehicles 1 and 3 assigned to distributors in cluster 1, Vehicle 2 assigned to distributor in cluster 3, Vehicles 4 assigned to distributor in cluster 2.

With the above structure for encoding solution vectors in the optimization problem under discussion, we can proceed to explain how the fitness of each solution is evaluated. In the proposed framework, the efficiency of the deployment pattern and resource allocation to distributors is assessed based on the incurred cost on the logistics distribution network's final cost. In a logistics distribution network, the system costs can be composed of three major components:

A) Cost of transporting products from the manufacturer to the distributor: This criterion calculates the costs required to transport products from the manufacturer to each of the distributors assigned to the sales clusters. Each transport vehicle has limited capacity, and the cost of transporting products by each vehicle is proportional to the distance traveled and its capacity. This criterion can be formulated as follows:

$$Cost_{MD} = \sum_{i=1}^{|D|} d_{M,D_i} \times \left\lceil \frac{y_{D_i}}{T_i} \right\rceil \times C_{avg} \tag{4}$$

Where in the above equation, $|D|$ specifies the number of distributors determined in the solution vector (the number of 1 bits in the first part of the solution vector), d_{M,D_i} represents the distances (in kilometers) between the manufacturer and distributor i, y_{D_i} denotes the total predicted sales for the cluster corresponding to distributor i, and T_i represents the capacity assigned to this distributor's vehicle. Finally, C_{avg} indicates the average cost of transporting products (fuel, wear and tear, and charges) for the vehicle assigned to this distributor per kilometer.

B) Cost of transporting products from the distributor to the retailer: This criterion represents the cost required to transport products from the distributors to the retailers. Since calculating this criterion for each retailer would be time-consuming, in the proposed optimization model, the distance from the distributor to the corresponding cluster center is considered. This criterion can be formulated as follows:

$$Cost_{DS} = \sum_{i=1}^{|D|} \sum_{j=1}^{K} \delta_{i,j} \times d_{D_i,c_j} \times \left\lceil \frac{y_{D_i}}{T_i} \right\rceil \times C_{avg} \tag{5}$$

In the above equation, K represents the number of sales clusters, and d_{D_i,c_j} indicates the distances (in kilometers) between distributor i and cluster center j. Additionally, $\delta_{i,j}$ is a binary function that equals 1 if distributor i is located in cluster j, and 0 otherwise.

C) Fixed costs of distributor deployment: This criterion describes the total fixed costs that must be incurred to establish and maintain the distributors determined in each solution vector and is denoted as $Cost_P$.

Considering the above cost factors, the total cost criterion in the distribution network can be formulated as the following evaluation function:

$$fitness = \alpha \times (Cost_{MD} + Cost_{DS}) + (1 - \alpha) \times Cost_P \tag{6}$$

In the above equation, $Cost_{MD}$ and $Cost_{DS}$ represent the costs of product transportation, calculated through Eqs. 4 and 5, respectively. The $Cost_P$ criterion indicates the total costs of deploying the distributors determined in the solution vector. Finally, α is the coefficient that represents the importance of variable costs (product transportation) compared to fixed costs (distributor deployment) and can be assigned a value in the range [0, 1]. The optimization problem modeled in this research includes only two main constraints:

1. In each sales cluster, at least one distributor must be stationed.
2. Each stationed distributor should be allocated one product transportation vehicle.

Input: population size (P), number of iterations (G), parameters β_0, β_1, and β_2
Output: Solution X indicating location of distributors and the allocation of resources to them

Step 1: The initial population is randomly generated based on the defined bounds for each optimization variable.
Step 2: The fitness of each solution (Capuchin) is calculated based on equation (7).
Step 3: The initial speed of each Capuchin agent is adjusted.
Step 4: Half of the Capuchin population is randomly selected as leaders, and the rest are determined as follower Capuchins.
Step 5: If the number of algorithm iterations reaches the maximum value G, proceed to step 13; otherwise, repeat the following steps:
Step 6: The CapSA lifespan parameter is calculated as follows [32]:

$$\tau = \beta_0 e^{\left(-\frac{\beta_1 g}{G}\right)^{\beta_2}} \quad (7)$$

In the above equation, g represents the current iteration count, and the parameters β_0, β_1, and β_2 have values of 2, 21, and 2, respectively.
Step 7: For each Capuchin agent (leader and followers) such as i, repeat the following steps:
Step 8: If i is a leader Capuchin, update its velocity based on the following equation [32]:

$$v_j^i = \rho v_j^i + \tau a_1 \left(x_{best_j}^i - x_j^i\right) r_1 + \tau a_2 \left(F - x_j^i\right) r_2 \quad (8)$$

In the above equation, index j represents the dimensions of the problem, and v_j^i denotes the velocity of Capuchin i in dimension j. x_j^i specifies the position of Capuchin i for variable j, and $x_{best_j}^i$ describes the best position of Capuchin i for variable j from the beginning until now. Additionally, r_1 and r_2 are two random numbers in the range [0,1]. Finally, ρ is the parameter that determines the influence of the previous velocity, with a value of 0.7.
Step 9: Update the new position of leader Capuchins based on their velocity and movement pattern.
Step 10: Update the new position of follower Capuchins based on their own velocity and the leader's position.
Step 11: Calculate the fitness of population members based on equation (7).
Step 12: If the overall population position has been updated, go back to step 5; otherwise, repeat the algorithm from step 7.
Step 13: Return the solution X with the minimum fitting as the optimal values for the thresholds T_s and T_R.

Algorithm 1. Optimizing problem using CapSA

Step 1: The initial population is randomly generated based on the defined bounds for each optimization variable. Step 2: The fitness of each solution (Capuchin) is calculated based on Eq. (7). Step 3: The initial speed of each Capuchin agent is adjusted. Step 4: Half of the Capuchin population is randomly selected as leaders, and the rest are determined as follower Capuchins. Step 5: If the number of algorithm iterations reaches the maximum value G, proceed to step 13; otherwise, repeat the following steps: Step 6: The CapSA lifespan parameter is calculated as follows[32]: $\tau = \beta_0 e^{\left(-\frac{\beta_1 g}{G}\right)^{\beta_2}}$ (8)In the above equation, g represents the current iteration count, and the parameters β_0, β_1, and β_2 have values of 2, 21, and 2, respectively.Step 7: For each Capuchin agent (leader and followers) such as i, repeat the following steps: Step 8: If i is a leader Capuchin, update its velocity based on the following equation [32]: $v_j^i = \rho v_j^i + \tau a_1 \left(x_{best_j}^i - x_j^i\right) r_1 + \tau a_2 \left(F - x_j^i\right) r_2$ (9)In the above equation, index j represents the dimensions of the problem, and v_j^i denotes the velocity of Capuchin i in dimension j. x_j^i specifies the position of Capuchin i for variable j, and $x_{best_j}^i$ describes the best position of Capuchin i for variable j from the beginning until now. Additionally, r_1 and r_2 are two random numbers in the range [0,1]. Finally, ρ is the parameter that determines the influence of the previous velocity, with a value of 0.7.Step 9: Update the new position of leader Capuchins based on their velocity and movement pattern.Step 10: Update the new position of follower Capuchins based on their own velocity and the leader's position.Step 11: Calculate the fitness of population members based on Eq. (7).Step 12: If the overall population position has been updated, go back to step 5; otherwise, repeat the algorithm from step 7.Step 13: Return the solution X with the minimum fitting as the optimal values for the thresholds T_s and T_R.

Experimental results

The model presented in this paper, was implemented using MATLAB 2020a. Our implementation is done in two phases. In the first phase, which tries to forecast the quantity of sales made by sellers, we compare the prediction error of the proposed method with that of different alternative methods. In the second phase, we have

the efficiency of the resource allocation technique, which we were able to minimize costs and boost resource efficiency in the product distribution network by using the proposed CapSA algorithm.

Phase 1: the effectiveness of the proposed approach in estimating the volume of sales

We examined the efficiency of the techniques used in the proposed method by considering two scenarios: Proposed (KMeans + GPR) and proposed (Single GPR).

- Proposed (KMeans + GPR) is related to the case where we initially perform clustering on sellers and when we get the clusters of sellers, we use a separate GPR model to predict the sales amount of each cluster.
- Proposed (Single GPR) refers to the case where we use a single GPR model to predict all sellers.

Initially, we explored the use of two well-established machine learning algorithms for sales prediction: Adaptive Neuro-Fuzzy Inference System (ANFIS) and Classification and Regression Tree (CART). We evaluated the performance of both ANFIS and CART models using historical sales data from each seller cluster. However, after careful consideration, we opted to implement GPR for the final sales prediction model within our framework. Non-linearity, uncertainty quantification, and computational efficiency were the main advantages that led to choosing GPR in the proposed method. Also, in addition to above cases, Leung et al.[21], Cai et al.[22], were considered for comparing with the proposed method. The experiments were performed using a 10-fold cross-validation method.

Reduced prediction error is shown to be the ideal condition for this measure in Fig. 4a, and this finding is confirmed by examining variations in RMSE. Every iteration in Fig. 4 corresponds to a single cross-validation fold. Ten folds are created from the whole dataset, and the model is trained on all but one of the folds for testing. By repeating this procedure for every fold, the model can be assessed on data that hasn't been seen yet and its performance on fresh data may be estimated with greater confidence.

Figure 4 illustrates how the suggested approach's RMSE trended downward and steadily over the course of the iterations. This means that for every cross-validation fold, the GPR model is constantly delivering good results on unknown data. These results indicate a higher likelihood of correct outputs from the proposed approach due to the reduced degree of prediction error and narrower range of error variations. The average value of 5.5 in the findings indicates that we were able to have a high sales forecast rate using the proposed method in comparison to other ways by clustering the sellers.

Compact box plots are used in Fig. 4b to display the average RMSE values across all repetitions of the change interval. The blue line in these graphs represents the range of RMSE changes over different iterations, while the gray line represents the average value of RMSE throughout all iterations. The mean of the mistake is indicated by the white dot inside each bar.

In Fig. 5 (a, b), we evaluated the percentage of the absolute value of the error in various repetitions and found that, in comparison to previous comparable methods, our proposed method had a smaller error with the minimum and maximum values of 0.14 and 0.19, respectively. MAPE can be expressed using the equation[33]:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i} \tag{9}$$

Where A_i Is the actual value, F_i is the forecast value, n is total number of observation.

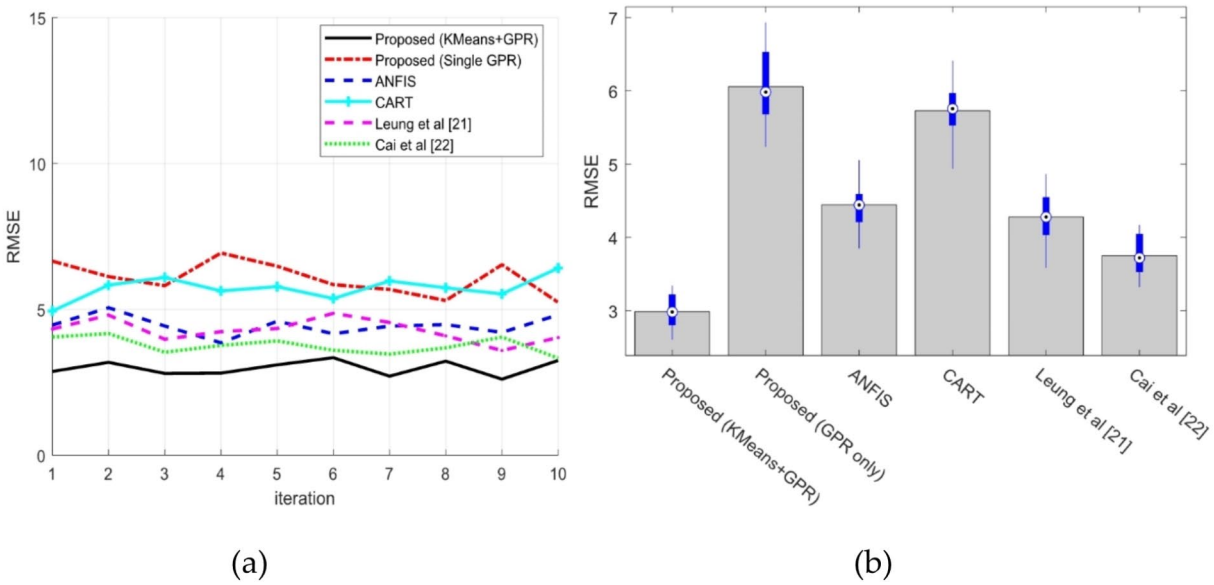

Fig. 4. The amount of RMSE in sales prediction.

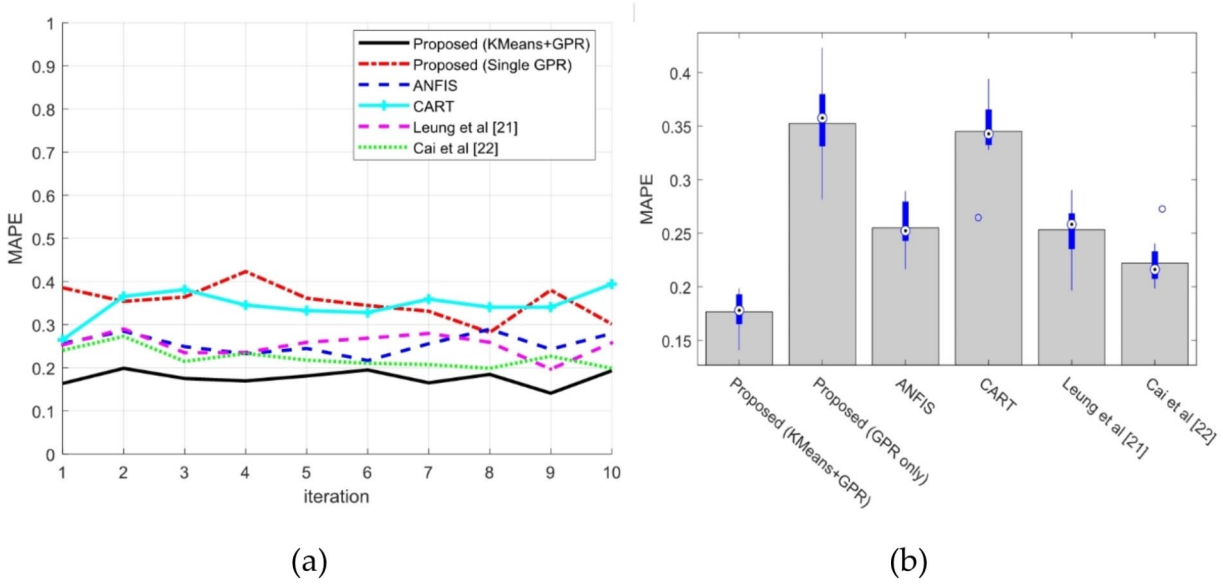

Fig. 5. The amount of MAPE in sales prediction.

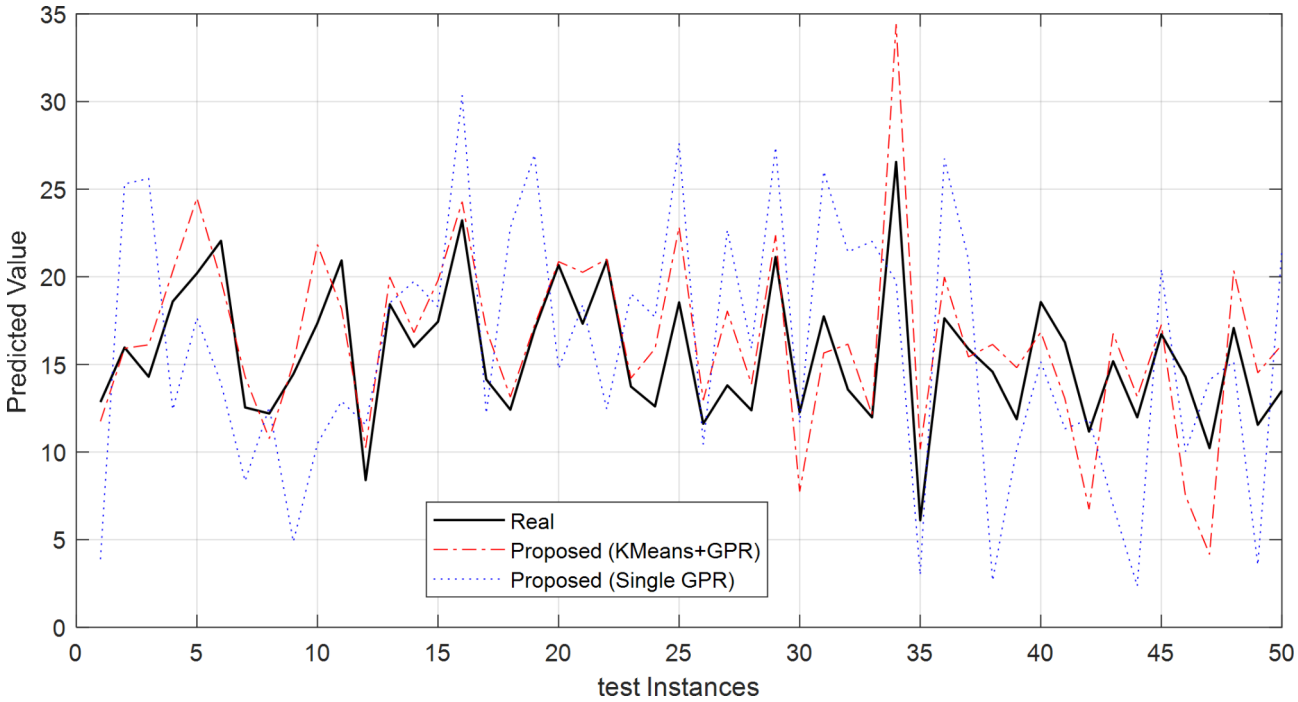

Fig. 6. Predicted value of real and proposed method.

The impact of clustering on the accuracy of sales forecasting is examined in Fig. 6. The black line in this diagram represents real sales, the blue dotted line represents the situation in which clustering is not done and only one GPR is utilized for forecasting, and the red dashed line represents the suggested strategy that combines K-mean and employs a GPR for each cluster, separately. We were able to develop a model that is better at estimating the volume of sales by using the clustering technique. Since the values predicted by the suggested approach are closer to the actual state, the results shown in this graph represent those values.

Despite the fact that R2 is an error-based measure, we compared four correlation measures (R^2, PLCC, SROCC, and CCC) for various approaches in Fig. 7. Since our suggested solution was able to satisfy both of these requirements at once, the greater the correlation criteria, and since R^2 is based on error, the lower it is, the better. The higher value of the correlation criteria demonstrates that the proposed method's predicted values are more

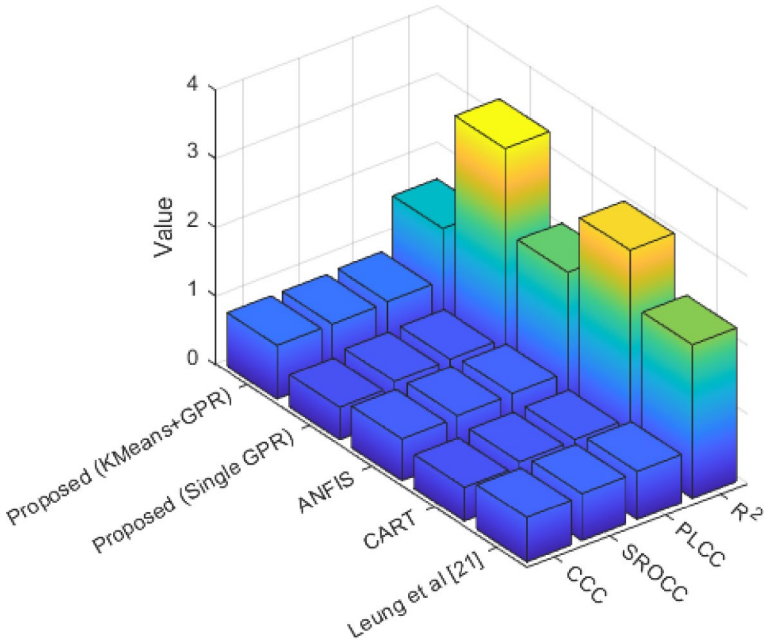

Fig. 7. Correlation measures of sales prediction.

Methods	RMSE	MAPE	*R* Squared	PLCC	SROCC	CCC
Proposed (KMeans + GPR)	2.9889	0.1766	0.6453	0.8033	0.7834	0.7837
Proposed (Single GPR)	6.0592	0.3526	0.3122	0.5588	0.5416	0.4715
ANFIS	4.4469	0.2552	0.4207	0.6486	0.6329	0.6055
CART	5.7286	0.3451	0.3274	0.5722	0.5580	0.4947
Leung et al.[21]	4.2800	0.2534	0.4931	0.7022	0.6866	0.6489
Cai et al.[22]	3.7544	0.2221	0.5266	0.7257	0.7154	0.6927

Table 3. Values of sales forecast correlation criteria.

in line with the actual sales values, which leads to less error as a result of the proposed method's more accurate modelling of the sales changes that occur in the real state. The R^2 criterion is clearly visible, demonstrating that the proposed approach has considerably lower error rates than competing approaches. Table 3 compares the correlation criterion, and it reveals that the RMSE and MAPE errors for our suggested approach are the lowest. Correlation criteria can be expressed using the equations[34]:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{10}$$

Where, r = Pearson correlation Coefficient.

x_i= x variable samples, y_i= y variable sample, $\bar{x}$= mean of values in x variable, $\bar{y}$= mean of values in y variable[33].

$$R^2 = \frac{SSR}{SST} \tag{11}$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2 \tag{12}$$

$$SST = \sum_i (y_i - \bar{y})^2 \tag{13}$$

- Where, SSR is sum of Squared Regression also known as variation explained by the model.
- SST is Total variation in the data also known as sum of squared total.
- y_i is the y value for observation i.
- $\bar{y}$ is the mean of y value.
- $\hat{y}_i$ is predicted value of y for observation i.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{14}$$

Where, ρ =Spearman's rank correlation coefficient.
d_i =difference between the two ranks of each observation.
n = number of observations.

Phase 2: performance of the proposed method in terms of resource allocation

In phase 2, we examined the costs and waste of resources and compared them to other comparative approaches.

Figure 8 shows the fitness changes in different iterations. The parameters we used for the Capuchins algorithm include the number of repetitions equal to 500 and the population size equal to 200, and we optimized the model according to these parameters. This diagram shows 10 changes in the optimization algorithm's fit in different iterations, which includes 2 parts:

1) in each Iteration, it shows the value of the best fitness.
2) It shows us the average fitness value of the population in each iteration. This graph shows that our proposed method is able to consistently reduce the fitness level in different iterations.

And it shows that we were able to get answers through the CapSA algorithm, the cost of which is continuously decreasing. The mean plot is continuously decreasing and tends towards better fitness, that is, our population is moving towards the global optimum, which indicates that the performance of this algorithm is correct and not stuck at the local optimum.

In Fig. 9, it shows us different cost criteria based on the number of clusters. In this diagram, we have set the horizontal axis based on the number of clusters and divided the data based on several clusters, and the purpose of this diagram and diagram 10 is to specify The optimal number of clusters is that we can get the best answer

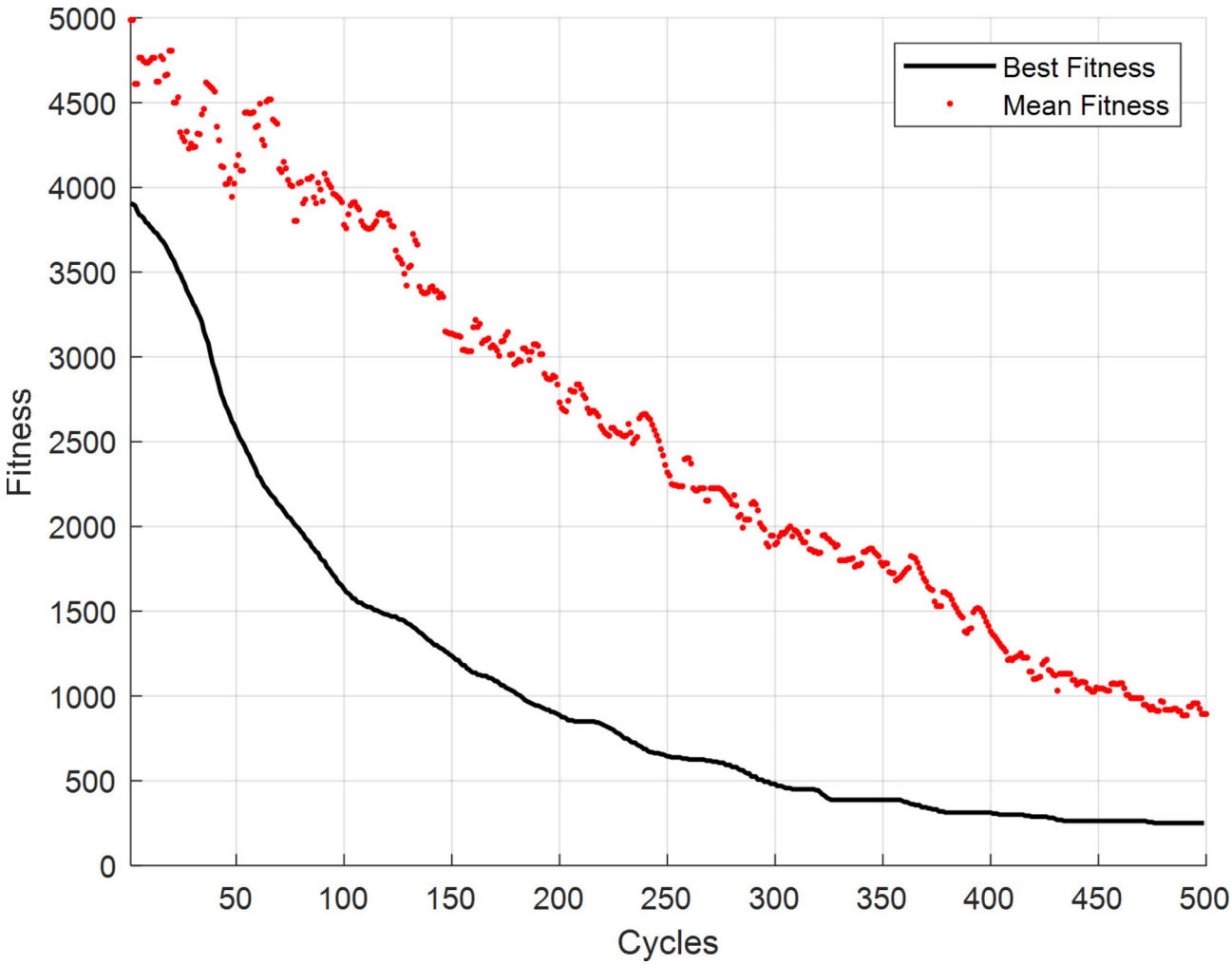

Fig. 8. Best and mean fitness values.

for the problem. In this diagram, we compared 3 cost criteria ($Cost_{MD}$, $Cost_{Ds}$, and $Cost_{p}$) that we had in relation to fitness.

$Cost_{p}$, every time we increase the number of clusters, we need a large number of distributors, which increases our fixed cost. $Cost_{MD}$ When we increase the number of distributors, we have to make more transfers between the manufacturer and the distributor, these transfers will increase and as a result, the necessary cost to transfer the manufactured products from the manufacturer to the distributor will increase. But $Cost_{Ds}$ decreases, the reason is that when we increase the number of clusters, we have many distributors available, whose distance and access cost to retail sellers is less. Total cost specifies the total cost, which shows us the sum of 3 costs for the number of different clusters.

In Fig. 10, we determined the optimal value of the clusters, according to Eq. 7, the α value determines how effective each of the distribution costs and relocation costs are in the fitness function. In this graph, we compared the value of total cost with different values of α. The higher the value of α, the depth of the graph moves to higher K's. The reason is that the more we increase α, the more important the moving costs become, and the less important the fixed cost, so it causes The algorithm tries to choose more K values to optimize the problem, so we chose the value of 9, which is the optimal state for us.

In Fig. 11, we have the amount of cost reduction and resource wastage, the higher the amount of resource reduction and the less resource wastage, the better. That we have compared the proposed method i.e. CapSA algorithm with the two methods of Wang et al.[25] and Redriguez et al.[18]. In Redriguez's method, he used linear programming algorithm to determine the location of the distributor. And in Wang's method, the authors try to specify routes for the producer to the distributor through optimization. In the proposed method, we were able to obtain two criteria (reduction of cost and waste of resources) at the same time, the reason is the use of Capuchin search algorithm to optimize the problem based on the amount of cost, which caused the cost to decrease compared to the comparison methods.

While the proposed optimization model prioritizes cost reduction, it inherently contributes to resource wastage reduction through optimized transportation planning. Minimizing transportation costs, a significant component of the fitness function (Eq. 7), directly translates to:

- Reduced vehicle usage: Optimized distributor placement and resource allocation minimize unnecessary travel distances, leading to fewer vehicle trips and decreased fuel consumption. This directly reduces wasted resources like fuel and associated emissions.
- Improved efficiency: The model promotes efficient utilization of vehicles by strategically assigning them to clusters based on predicted sales volume. This reduces the need for additional vehicles or inefficiently loaded trips, further minimizing resource wastage.

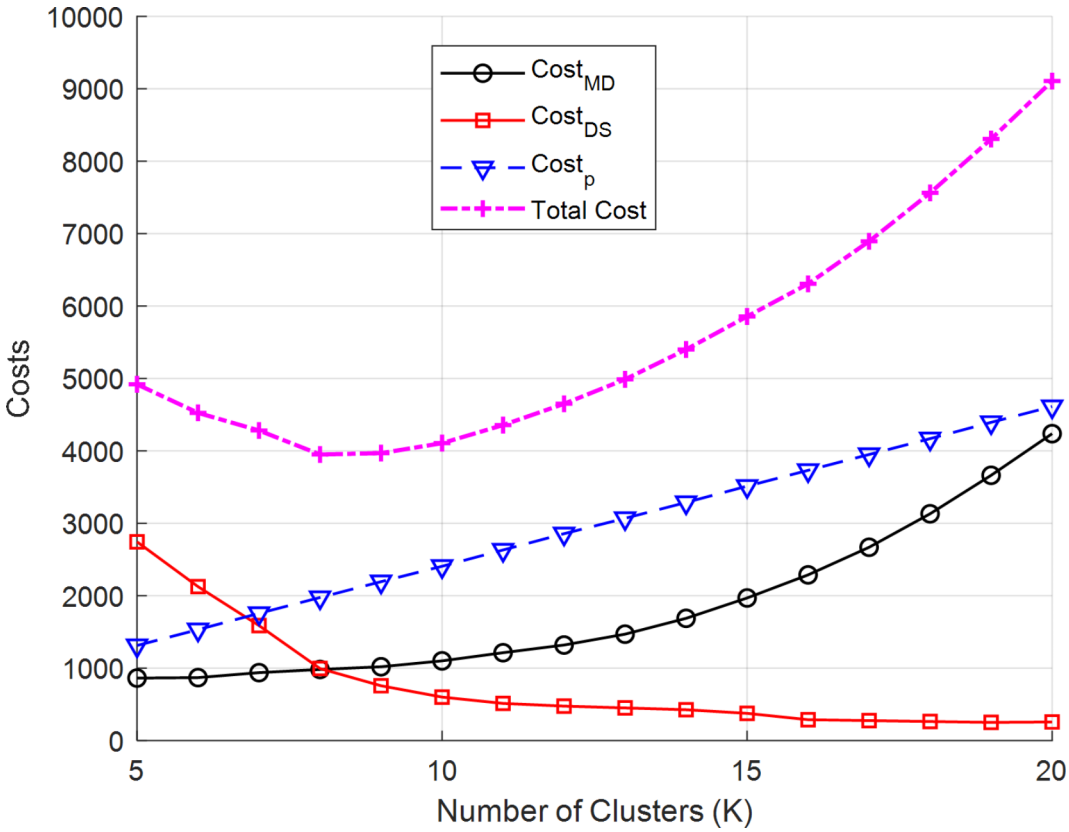

Fig. 9. Cost amounts.

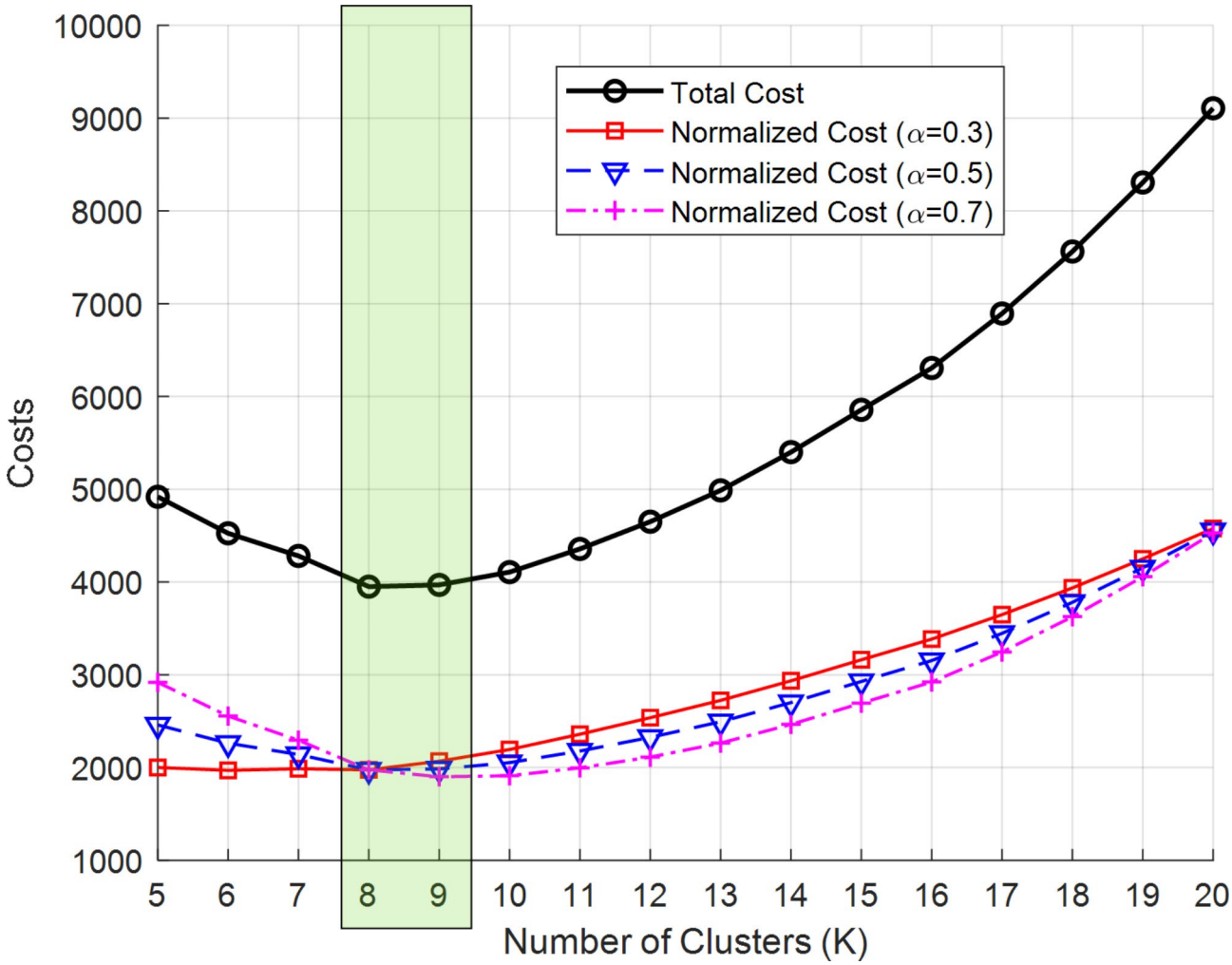

Fig. 10. Values of costs by calculating alpha.

Furthermore, our approach indirectly impacts resource utilization through improved logistics efficiency:

- Reduced packaging waste: Fewer and more efficient trips could potentially lead to less packaging material used for product transportation, contributing to resource conservation.
- Lower maintenance needs: Optimized vehicle usage can translate to reduced wear and tear, requiring less frequent maintenance and associated resource consumption.
- Energy savings: Streamlined distribution networks with fewer trips may contribute to lower energy consumption at distribution centers, further reducing resource wastage.

While our primary focus is cost minimization, these secondary effects highlight the broader positive impact of the proposed method on resource utilization.

The findings demonstrate that we have cost reduction with maximum value and resource reduction with minimum value, yielding values of 34.76 and 15.6, respectively, demonstrating the efficacy of our proposed approach.

Managerial insights

Adopting the suggested tripartite methodology provides noteworthy benefits for professionals in the cold chain logistics sector overseeing frozen food distribution systems. The following are some significant managerial insights from this study:

- Substantial Cost Reduction: Our methodology has the ability to reduce costs for logistics organizations by 34.76% compared to existing systems. This reduction is obtained through optimizing distributor placement and resource allocation based on precise sales estimates and a combined assessment of holding and transportation costs.
- Enhanced Resource Efficiency: Based on anticipated demand fluctuations, the data-driven strategy enables a more intelligent distribution of resources among various seller clusters, resulting in a 15.6% reduction in

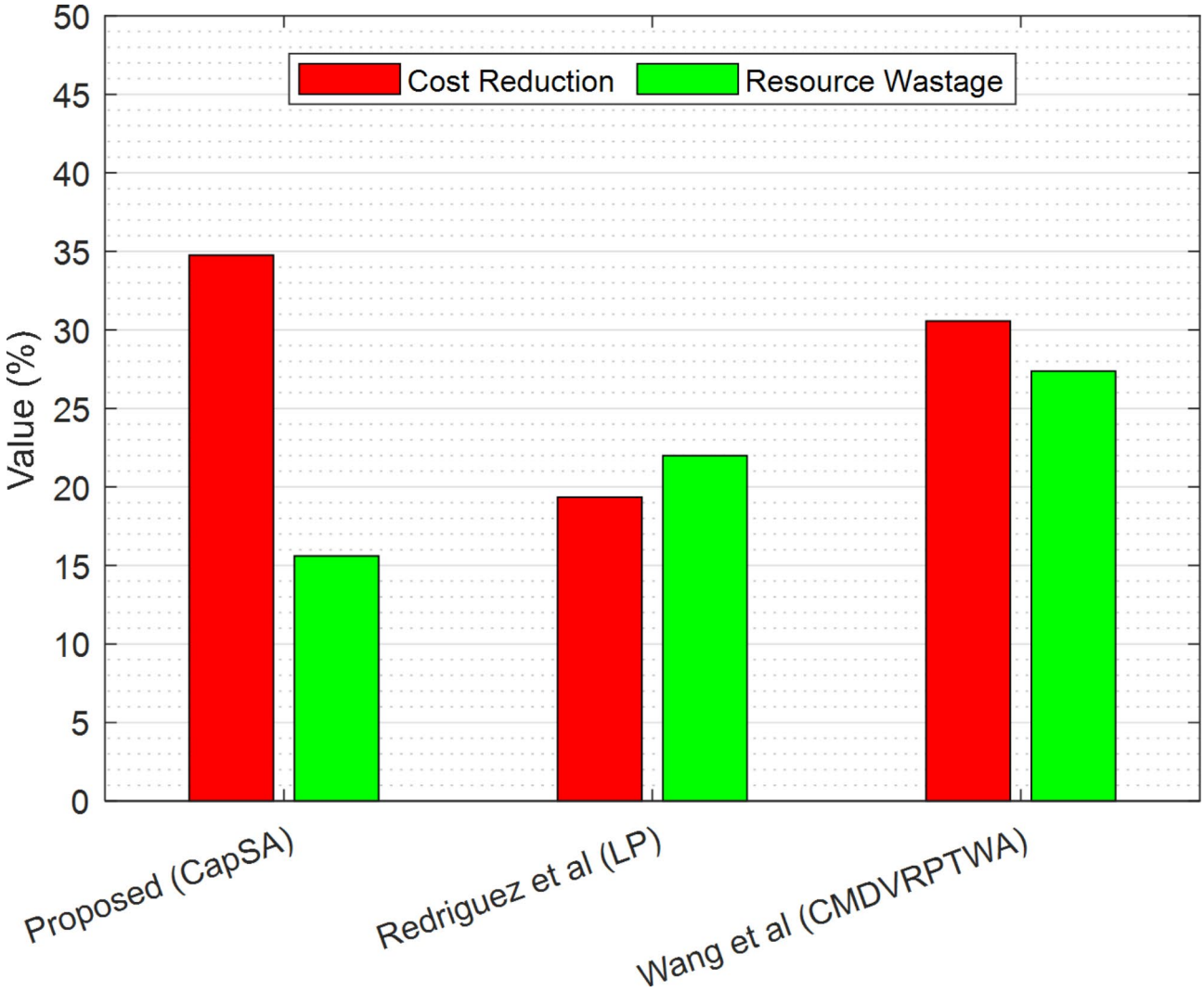

Fig. 11. The amount of cost reduction and waste of resources.

resource wastage. This may lead to increased effectiveness and decreased waste of resources, such as storage space and transportation vehicles.

- Improved Decision-Making: For vendors who are geographically concentrated, the model offers insightful information about future sales trends and provides predictions with average MAPE of 0.17. Logistics managers can use this information to make better-informed decisions about workforce numbers, inventory control, and general network optimization techniques.

In Summary, the proposed methodology can be highly useful for cold chain logistics managers to optimize the frozen food supply chain networks and to realize substantial cost reductions, increased resource utilization and improved supply chain performance. With the help of data analysis and logics optimization, the professional of logistic can make the right decision and can manage the logistic flow according to the requirement of cold chain industry.

Conclusion

This research investigated the optimization of frozen goods distribution networks using a machine learning-based approach. The proposed method, consisting of seller clustering, sales volume prediction with GPR, and distributor placement/resource allocation with the Capuchin Search algorithm, aimed to achieve optimal distributor positioning and minimize costs. The key findings and their significance are highlighted below:

- Effective sales volume prediction: The GPR models achieved an average RMSE error of 2.98 and MAPE error of 0.17, demonstrating successful capture of sales patterns and accurate forecasting within each seller cluster.
- Significant cost reduction: Compared to existing methods, the proposed approach reduced costs by a notable percentage of 34.76%. This translates to savings for businesses involved in frozen goods distribution.

- Reduced resource wastage: The method achieved a percentage reduction in resource wastage 15.6%, improving overall resource utilization and minimizing unnecessary expenses.
- Distribution network optimization: By simultaneously considering transportation and holding costs, the Capuchin Search algorithm effectively balanced competing objectives, leading to a globally optimal solution for distributor placement and resource allocation.

The proposed method offers several advantages over existing approaches:

- Data-driven approach: GPR relies on historical sales data to learn patterns and make predictions, leading to more accurate and adaptable sales forecasts compared to static methods.
- Model optimization: Simultaneous consideration of cost and resource usage ensures efficient resource allocation and cost savings.
- Flexibility: The framework can be adapted to different network sizes and data characteristics, making it applicable to various frozen goods distribution scenarios.

Future research directions include:

- Exploring alternative clustering algorithms for potentially better seller grouping.
- Investigating techniques for incorporating additional factors, such as product type or weather conditions, into the sales volume prediction models.
- Developing a dynamic optimization framework that can adapt to changes in network conditions or demand patterns.

In conclusion, this study presented a novel and effective machine learning-based approach for optimizing frozen goods distribution networks. The method demonstrated significant cost reduction, resource wastage reduction, and accurate sales volume prediction, offering a valuable tool for businesses seeking to improve efficiency and profitability in their frozen goods distribution operations. Future research directions will explore further enhancements and expand the applicability of this approach.

Data availability

All data generated or analysed during this study are included in this published article.

Received: 2 November 2023; Accepted: 10 September 2024
Published online: 28 September 2024

References

1. Garside, A. K. An optimization model for cold chain food distribution. *Int. J. Res. Industrial Eng.* **8**(3), 243–253 (2019).
2. Huang, W., Wang, X., Zhang, J., Xia, J. & Zhang, X. Improvement of blueberry freshness prediction based on machine learning and multi-source sensing in the cold chain logistics. *Food Control.* **145**, 109496 (2023).
3. Yang, Z., Xu, J., Yang, L. & Zhang, X. Optimized dynamic monitoring and Quality Management System for Post-harvest Matsutake of different preservation packaging in Cold Chain. *Foods*, **11**(17). (2022).
4. Qin, G., Tao, F. & Li, L. A vehicle routing optimization problem for cold chain logistics considering customer satisfaction and carbon emissions. *Int. J. Environ. Res. Public Health.* **16** (4), 576 (2019).
5. Xu, X. & Wei, Z. Dynamic pickup and delivery problem with transshipments and LIFO constraints. *Comput. Ind. Eng.* **175**, 108835 (2023).
6. Liu, G., Hu, J., Yang, Y., Xia, S. & Lim, M. K. *Vehicle routing problem in cold Chain logistics: A joint distribution model with carbon trading mechanisms* Vol. 156, 104715 (Resources, Conservation and Recycling, 2020).
7. Li, Y., Lim, M. K. & Tseng, M. L. A green vehicle routing model based on modified particle swarm optimization for cold chain logistics. *Industrial Manage. Data Syst.* **119**(3), 473–494 (2019).
8. Wang, S., Tao, F., Shi, Y. & Wen, H. Optimization of vehicle routing problem with time windows for cold chain logistics based on carbon tax. *Sustainability* **9**(5), 694 (2017).
9. Wang, Z. et al. Blockchain-based framework for improving supply chain traceability and information sharing in precast construction. *Autom. Constr.* **111**, 103063 (2020).
10. Dai, J., Che, W., Lim, J. J. & Shou, Y. Service innovation of cold chain logistics service providers: a multiple-case study in China. *Ind. Mark. Manage.* **89**, 143–156 (2020).
11. Ferrentino, R. & Boniello, C. Customer satisfaction: a mathematical framework for its analysis and its measurement. *Comput. Manage. Sci.* **17**, 23–45 (2020).
12. de Aquino, J. T., de Melo, F. J. C., Jeronimo, T. D. B. & de Medeiros, D. D. Evaluation of quality in public transport services: the use of quality dimensions as an input for fuzzy TOPSIS. *Int. J. Fuzzy Syst.* **21**, 176–193 (2019).
13. Kumar, S. N. & Panneerselvam, R. A survey on the vehicle routing problem and its variants. (2012).
14. Xu, X., Lin, Z., Li, X., Shang, C. & Shen, Q. Multi-objective robust optimisation model for MDVRPLS in refined oil distribution. *Int. J. Prod. Res.* **60**(22), 6772–6792 (2022).
15. Pitaloka, D. A. & Mahmudy, W. F. PENYELESAIAN VEHICLE ROUTING PROBLEM WITH TIME WINDOWS (VRPTW) MENGGUNAKAN ALGORITMA GENETIKA HYBRID. *J. Environ. Eng. Sustainable Technol.* **1**(2), 104–110 (2014).
16. Zhang, B. The Optimization of Distribution Path of Fresh Cold Chain Logistics Based on Genetic Algorithm. Computational Intelligence and Neuroscience, 2022. (2022).
17. Gámez-Albán, H. M. & Mejía-Argueta, C. León Espinosa de los Monteros. *Ingeniare Revista Chil. de ingeniería*. **25** (4), 619–632 (2017). Diseño de una red de distribución a través de un modelo de optimización considerando agotados.
18. Rodríguez, J. V., Niño, J. P. C., Negrete, K. A. P., Mercado, D. C. & Fontalvo, L. A. Optimization of the distribution logistics network: a case study of the metalworking industry in Colombia. *Procedia Comput. Sci.* **198**, 524–529 (2022).
19. Ariafar, S., Ahmed, S., Choudhury, I. A. & Bakar, M. A. Application of fuzzy optimization to production-distribution planning in supply chain management. Mathematical Problems in Engineering, 2014. (2014).
20. Li, G. Development of cold chain logistics transportation system based on 5G network and internet of things system. *Microprocess. Microsyst.* **80**, 103565 (2021).

21. Leung, K. H., Mo, D. Y., Ho, G. T., Wu, C. H. & Huang, G. Q. Modelling near-real-time order arrival demand in e-commerce context: a machine learning predictive methodology. *Industrial Manage. Data Syst.* **120**(6), 1149–1174 (2020).
22. Cai, W., Song, Y. & Wei, Z. Multimodal data guided spatial feature fusion and grouping strategy for E-commerce commodity demand forecasting. *Mob. Inform. Syst.* **2021**, 1–14 (2021).
23. Chen, Y. H. Intelligent algorithms for cold chain logistics distribution optimization based on big data cloud computing analysis. *J. Cloud Comput.* **9**, 1–12 (2020).
24. Lim, M. K., Li, Y. & Song, X. Exploring customer satisfaction in cold chain logistics using a text mining approach. *Industrial Manage. Data Syst.* **121**(12), 2426–2449 (2021).
25. Wang, Y. et al. Collaborative multi-depot logistics network design with time window assignment. *Expert Syst. Appl.* **140**, 112910 (2020).
26. Lin, T. X. & Wu Zh, Pan, W. T. Optimal location of logistics distribution centres with swarm intelligent clustering algorithms. *PLOS ONE*. **17** (8), e0271928 (2022).
27. Tordecilla, R. D., Juan, A. A., Montoya-Torres, J. R., Quintero-Araujo, C. L. & Panadero, J. Simulation-optimization methods for designing and assessing resilient supply chain networks under uncertainty scenarios: a review. *Simul. Model. Pract. Theory*. **106**, 102166 (2021).
28. Nayeri, S., Paydar, M. M., Asadi-Gangraj, E. & Emami, S. Multi-objective fuzzy robust optimization approach to sustainable closed-loop supply chain network design. *Comput. Ind. Eng.* **148**, 106716 (2020).
29. Hasani, A., Mokhtari, H. & Fattahi, M. A multi-objective optimization approach for green and resilient supply chain network design: a real-life case study. *J. Clean. Prod.* **278**, 123199 (2021).
30. Nagurney, A. Optimization of supply chain networks with inclusion of labor: applications to COVID-19 pandemic disruptions. *Int. J. Prod. Econ.* **235**, 108080 (2021).
31. Swiler, L. P., Gulian, M., Frankel, A. L., Safta, C. & Jakeman, J. D. A survey of constrained gaussian process regression: approaches and implementation challenges. *J. Mach. Learn. Model. Comput.*, **1**(2). (2020).
32. Braik, M., Sheta, A. & Al-Hiary, H. A novel meta-heuristic search algorithm for solving optimization problems: capuchin search algorithm. *Neural Comput. Appl.* **33**, 2515–2547 (2021).
33. Zeiml, S., Seiler, U., Altendorfer, K. & Felberbauer, T. Simulation evaluation of automated forecast error correction based on mean percentage error. In 2020 Winter Simulation Conference (WSC) (pp. 1572–1583). IEEE. (2020), December.
34. Yang, Q. et al. Linear correlation analysis of ammunition storage environment based on Pearson correlation analysis. In Journal of physics: Conference series (Vol. 1948, No. 1, p. 012064). IOP Publishing. (2021), June.

Author contributions
All authors wrote the main manuscript text. All authors reviewed the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to J.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.