Solving Multiple Distribution Center Location Allocation Problem Using K-Means Algorithm and Center of Gravity Method
Take Jinjiang District of Chengdu as an example

Chunling Cai\(^1\), Yong Luo\(^2\)\(^*,\) Yuehong Cui\(^3\), Feng Chen\(^4\)

\(^1\)College of Information Engineering Sichuan Agricultural University Ya’an, China
\(^2\)School of Economics and Management Sichuan Tourism University Chengdu 610100, China
\(^3\)College of Information Engineering Sichuan Agricultural University Ya’an, China
\(^4\)College of Information Engineering Sichuan Agricultural University Ya’an, China

\(^*\)Corresponding author: luoyongcds@163.com
\(^1\)17608019254@163.com
\(^2\)1356373432@qq.com, \(^4\)381162109@qq.com

Abstract— In the logistics system, logistics nodes play an important role. The centre-of-gravity method is one of the most basic methods in solving the problem of single facility location. The thesis first uses the K-means clustering algorithm to divide the demand points into k clusters, and the k value is selected by the elbow method. Then, the k clusters are regarded as a single facility location problem and analyzed by the center of gravity method. In practical applications, the discrete model analysis can be cited to further obtain the optimal solution.

1. Introduction
In the logistics system, the logistics nodes play an important role in linking up and down. As an intermediate bridge linking the receiving and sending locations, the location of logistics nodes has an important impact on the economic growth and planning of the entire region. Reasonable location selection can help companies optimize the spatial layout of product sales, thereby enabling companies to reduce costs and improve profits, enhance comprehensive competitiveness. Therefore, the location and layout of the distribution center must take into account the characteristics of its own operations, product characteristics and traffic conditions, and select the location of the distribution center based on full investigation and analysis.

2. Data preprocessing

2.1. Selecting a Template
We select a total of 5,000 data items in the Jinjiang District of Chengdu's express delivery data for August to carry out research. The original data set contains 9 features such as the waybill number, the sending province, the sending address, and the receiving address. We select the feature of the recipient...
address and then uses the "XO map" developed and explored by GIS9 to convert the address to coordinates. Export the two data of longitude and latitude, and use two different methods to research and analyze the data set.

First, visualize the processed data, and the result is shown in Figure 1.

![Figure 1. Distribution of original demand points in Jinjiang District, Chengdu](image1)

3. Algorithm design

3.1. Flowchart

![Figure 2. Flowchart](image2)
3.2. Logistics location model
The Centre-of-gravity method is one of the most basic methods in solving the problem of single facility location. What the single facility location wants is to choose a suitable distribution center, so that the total transportation cost from the distribution center to each demand point is the least. The basis of the gravity center model is that the goal of single facility location is the minimum total transportation cost from the distribution center to each demand point. The formula is:

\[ x = \frac{\sum_{i=1}^{n} \text{Transport distance} \times \text{Demand}_i \times \text{X}_i}{\sum_{i=1}^{n} \text{Transport distance} \times \text{Demand}_i} \]  \tag{1}

\[ y = \frac{\sum_{i=1}^{n} \text{Transport distance} \times \text{Demand}_i \times \text{Y}_i}{\sum_{i=1}^{n} \text{Transport distance} \times \text{Demand}_i} \]  \tag{2}

K-means clustering algorithm has many advantages, such as simple principle, easy implementation, fast convergence speed, and better clustering effect on large data sets. According to the distribution characteristics of the actual express delivery network, the clustering algorithm is used for the optimal location problem to the distribution center. The location problem is converted into a problem of first classifying massive data and then solving the categories one by one. Based on the above analysis, the main steps of the algorithm design in this paper are: First, the K-means clustering algorithm is used to divide the demand points into k clusters, and then the k clusters are regarded as a single facility location problem and analyzed by the center of gravity method. In practical applications, the choice of k values can be specified by the decision maker, or the best k values can be found by the method of finding the SSE inflection point by the elbow method. The formula is:

\[ SSE = \sum_{i=1}^{k} \sum_{x \in C_i} (x - \mu_i)^2 \]  \tag{3}

\( C_i \) is the i-th cluster, \( x \) is the sample point in \( C_i \), \( \mu_i \) is the centroid of \( C_i \), SSE is the clustering error of all samples, and represents the degree of clustering effect.

In practical applications, when the center of gravity method is used to solve the optimal location center, due to the limited influence factors it can find. We can first use the center of gravity method to solve multiple distribution center candidate points, and then introduce a discrete model to select the most optimal location center [1]. The objective function of the discrete model can be composed of three items: total freight cost, variable costs incurred by managers’ operations, and fixed costs for building distribution centers. The mathematical expression formula is as follows:

\[ \text{min} F(x) = \alpha \left( \sum (A + B) X + \sum V G \right) + \beta J \]  \tag{4}

\( \alpha \) and \( \beta \) are weighting coefficients, and \( \alpha + \beta = 1 \), \( \alpha, \beta \in (0, 1) \); A is the unit freight rate for the delivery of goods from the production site to the distribution center; B is the unit freight rate for delivering goods from the distribution center to the demand point; \( X \) is the number of products delivered to the demand point through the distribution center at the production site; \( V \) is the unit variable cost incurred by the distribution center for product operation; \( G \) is the number of products passing through the distribution center; \( J \) is the fixed construction and operation cost of the distribution center;

The total freight cost is relatively independent of the variable costs incurred by the operation. The fixed costs in the construction and distribution are linearly complementary to the first two items. Therefore, the addition rule in the multi-objective decision analysis can be used to merge the first two costs with the third one, so that the total cost reaches the minimum.

3.3. Selecting location center
In choosing the number of logistics centers to be built, we chose the elbow method to solve the k values selection problem. With the increase of the number of clusters k, the sample division is gradually refined. When the aggregation within each cluster is gradually increased, the squared error SSE will gradually decrease. In addition, when the value of k is close to the true number of clusters, the gain of the degree of aggregation obtained by increasing k values will quickly decrease, and the decrease of SSE will decrease sharply at this time, and will tend to be gentle as the value of k continues to increase. The performance of the example data (1000 pieces) in Jinjiang District of Chengdu is shown in Figure 4, showing that the results have converged. Substituting the k values
found according to SSE into the K-means algorithm, the clustering effect between the two sets of data is shown in Figures 4 and 5.

![Figure 3. Relationship between the number of centers to be built and the sum of squared errors](image)

Observe the most appropriate k values by figure 3. It can be seen intuitively from Fig. 3 that when k=10, the curve has an inflection point and reaches the minimum value. When k>10, the mean value of the intra-class difference increases in a small range, which proves that the logistics network has achieved good coverage when establishing 10 distribution centers. With the increase of distribution centers, the coverage effect has decreased. That is, when cities are divided into 10 categories, the similarity between the categories is high and the clustering results are good.

Use this algorithm to visualize the processed data through R studio, as shown in Figure 4 and Figure 5. It can be seen that the k value sought by the elbow method has a good clustering effect on the Chengdu Jinjiang District sample data (1000 items) and the Chengdu Jinjiang District sample data (2000 items). The experiment was carried out on an i5-10210U laptop computer, and it took an average of 260 seconds. Therefore, this method is reasonable and efficient for addressing multiple distribution centers.

![Figure 4. The performance of K-means algorithm on the distribution data set of Chengdu express receipt points (1000)](image)
First, we divide the data set into multiple clusters using the K-means clustering algorithm, its shape is variable, the clustering speed is fast, and it can effectively deal with noise. It can be seen from the figure that the clustering effect on multiple data sets is good. On this basis, we treat each cluster as a single distribution center location and use the center of gravity method to solve

4. Conclusion
In the process of data preprocessing, we generated a data set named the distribution of express delivery points in Jinjiang District, Chengdu, which contains 16 features such as the detailed address, longitude, and latitude of the recipient.

We first observe the distribution of the analysis data (courier data for food in Jinjiang District of Chengdu), and then find the characteristics of the shape of the network distribution. We combine the simple and practical K-means clustering algorithm with the center of gravity method to simplify the multi-facility location problem into two steps. The first step is to use K-means clustering algorithm for cluster analysis to divide the data set into multiple clusters with arbitrary shapes. In the second step, the center of gravity method is used to solve each cluster. In this way, the location problem of multiple distribution centers has been changed to the location problem of multiple single distribution centers. From the final experimental results, the method is feasible.

ACKNOWLEDGMENT
Deepest gratitude to my supervisor, Miss Luo, who has walked me through all the stages of the writing of this thesis. During the whole writing stage, she gave me great help. With her patient and enlightening guidance, my thesis can reach the current level. I am also very grateful to all my teachers who have helped me develop fundamental and necessary academic abilities. At the same time, I am also very grateful to my classmates who have given me great help on the road to learning. They are the cutest people in my life.

Research on Location Selection and Logistics Mode Optimization of Agricultural Products Distribution Center approved by Social Science Federation of Sichuan Agricultural University (No.2019PTYB13).

Project from Strategic Research Center of Potato Main Grain which is a Key research base of Humanities and Social Sciences approved by Sichuan Provincial Department of Education (No.MLS1606).
REFERENCES

[1] Mao-Sheng Y, Hua J. Research on Location Selection of Distribution Center based on Gravity Method and Discrete Model[J]. Railway Transport and Economy, 2007.

[2] Hansen P, Mladenovic N. Variable neighborhood search for the p-median[J]. Journal of Heuristics, 2004, 10(3): 293-314.

[3] Mladenović N, Brimberg J, Hansen P, et al. The p-median problem: A survey of metaheuristic approaches[J]. European Journal of Operational Research, 2007, 179 (3): 927-939.

[4] Colmenar J M, Greistorfer P, Marti R, et al. Advanced greedy randomized adaptive search procedure for the obnoxious p median problem[J]. European Journal of Operational Research, 2016, 252 (2): 432-442.

[5] Sheu J B, Lin Y S. Hierarchical facility network planning model for global logistics network configurations[J]. Applied Mathematical Modelling, 2012, 36(7): 3053-3066.

[6] Park, S. & T. E. Lee & C. S Sung. A three-level supply chain network design model with risk-pooling and lead times[J]. Transportation Research Part E.2010.46:563-581.

[7] Dong K F, Gan H C, Zhang H Z. Distribution center location model based on economics and timeliness[J]. Journal of University of Shanghai for Science and Technology, 2013.