Research on C-end Customer Demand Based on Cluster Analysis

Zhenlin Wei, Lan Wang* and Yuzi Ouyang
School of transportation, Beijing Jiaotong University, Beijing, China
Email: 15510928656@163.com;

Abstract. With the rapid development of Internet economy, the problem of "last kilometer" distribution difficulty in B2C e-commerce has become one of the pain points of the whole industry, and improving the customer service level of C-end has become an important content of enterprise development. The change of C-end demand distribution will have a significant impact on the location of end-of-line service outlets, so this paper studies the spatial distribution of C-end customer demand points. According to the significant spatial clustering of C-end customer demand points in geographical space, K-means algorithm and DBSCAN algorithm are selected to cluster the demand points. The clustering results of the two algorithms are evaluated by the internal distance and the average distance between clusters. The comparison results show that k-means algorithm has better clustering effect on the demand points, and its clustering results can also be used as the end the main measurement index of service network layout provides services for terminal service network location.

Keywords. Logistics engineering, K-means algorithm, DBSCAN algorithm, Demand distribution.

1. Introduction
In recent years, the development of new retail is in full swing, with new formats and new scenes emerging, which makes the demand of C-end customers for express service and terminal real-time logistics continue to be strong. New retail aims to allocate real-time, efficient and cost-effective resources, and deeply tap customer demand to adapt to the change of circulation business model. Many new retail formats are also around the "logistics front" development, to improve customer satisfaction at the C end, to achieve the end of the distribution of instant and convenient purposes. At present, there is still no perfect urban end distribution system in China, and problems such as unreasonable spatial distribution of end service outlets and uneven distribution quality occur frequently, resulting in low efficiency of end distribution. It can be seen that the change of retail formats in the future will promote the efficiency and re combination of circulation links, so it is necessary to further study the needs of C-end customers.

As far as the current situation of e-commerce in China is concerned, B2C mode directly serves C-end customers. For the user demand of the smallest unit, it has a wide range of demand distribution, because the main body of C-end demand includes government personnel, office workers, students, households, etc. as long as there is population and economic activity in the region, there will be C-end distribution demand; its demand volume is small and frequency is large, because C-end distribution goods usually only meet the individual or family needs Daily demand, most of which are small daily necessities, clothes, documents, etc. the goods are usually light in weight, small in volume, scattered in purchase time, and need to be distributed in small batch and multiple frequency.
2. Solution Ideas
Because of the above spatial demand distribution characteristics, C-end customers are regarded as the most important resources of enterprises, and customer demand as the center becomes an effective method and tool to improve competitiveness. Using clustering algorithm to cluster the spatial data of C-terminal demand point, we can find the customer behavior mode, carry out customer zoning planning, and provide help for enterprises to locate the terminal service point.

Clustering algorithm can effectively divide the data set according to a certain standard, and classify the similar data into the same subset. The C-end customer needs studied in this paper are not only widely distributed in geographical space, but also show significant spatial aggregation. The clustering algorithm can divide the discrete demand into a group according to a certain criterion, realize the spatial division of the distribution area, and use the clustering center to represent the location of customer cluster points, so as to solve the problem that it is difficult to determine the demand points in the location of terminal service points. At present, there are many researches on the clustering algorithm of demand points at home and abroad. Li M et al. [1] selected two-stage K-means clustering algorithm to cluster demand points and realize the division of distribution areas. Lan Y H et al. [2] use clustering method to divide cities based on logistics service demand network. Zhu Xiangzhou [3] uses clustering algorithm to analyze the hot spot area of spatial data of intelligent scenic spot, designs experiments to compare K-means and DBSCAN clustering algorithm, and proves that DBSCAN algorithm has more advantages in hot spot area analysis. Wen H Y et al. [4] used the improved fuzzy clustering method to cluster the logistics park, and constructed the F index according to the internal distance and the distance between clusters to determine the optimal number of clusters. Bychko V A [5] proposes a clustering algorithm to automatically determine the optimal location of the cluster center in the logistics task. The cluster center is regarded as a temporary storage point, which is connected with other terminals through the shortest path. Zhou Xiang et al. [6] analyzed the distribution of customer points in online retail distribution, and designed a dynamic grid density clustering algorithm based on the distribution characteristics of customer points. Wang Pengfei [7] based on the k-means algorithm to cluster the potential customer groups, to find out which areas of the user purchase demand is relatively concentrated, and select a representative center point. Nam N M et al. [8] proposed a new clustering model based on the square distance of convex set, and used it in facility location and clustering. The above K-means and DBSCAN algorithms both look for clusters that use all attributes, that is, they do not look for clusters that may only involve a subset of attributes, and they are all partition clustering algorithms that assign each object to a single cluster.

In conclusion, this paper analyzes the spatial distribution characteristics of C-end customer demand, and uses K-means and DBSCAN clustering algorithm to aggregate the discrete customer points.

3. Model Establishment

3.1. Problem Description
Under the known C-end customer demand points, the geographical location of each demand point is known. Considering that the sample attribute in this paper is latitude and longitude, we can’t directly use the distance between coordinates to measure the similarity. Therefore, this paper uses the calculation formula of longitude and latitude distance to calculate the distance between two demand points, and takes it as the similarity measurement formula.

3.2. Symbol Description
Let $X = \{x_1, x_2, ..., x_n\}$ be the set of C-end customer demand points, $n$ be the number of C-end customer demand points, and sample $x_i = \{x_{i1}, x_{i2}, ..., x_{im}\}$ be described by $m$-attribute or characteristic $S = \{s_1, s_2, ..., s_m\}$. K-means clustering algorithm is to minimize the constrained nonconvex function $F$, that is, $X$ is divided into $K$ classes, so that the sum of squares from each
The sample point to the final clustering center is the smallest. \( w_{ij} \) is a binary variable, with a value of 0 or 1; \( W \) is the membership matrix between sample \( x_i \) and various types; \( m \) is the attribute dimension of the sample; \( s_i \) is the \( l \)-th attribute of the sample; \( z_j \) is the \( j \)-th center, and \( z_j = [z_{j1}, z_{j2}, \ldots, z_{jm}] \) consists of \( m \) components; \( Z \) is the clustering center matrix, with \( Z = [z_1, z_2, \ldots, z_k] \); \( d \) is the euclidean distance; \((a_i, b_i)\) is the longitude and latitude coordinates of the \( i \)-th demand point. The \( \varepsilon \)-neighborhood of any point in the sample set refers to the circle with the point as the center and \( \varepsilon \) as the radius; the density threshold is represented by \( \text{Minpts} \).

### 3.3. K-means Clustering Algorithm

K-means clustering algorithm is to judge the similarity between objects by distance size, that is to say, the points close to the sample set are clustered into a group. The main factors affecting the clustering results are to select the set \( X \) of \( n \) C-end customer demand point data and the number \( k \) of customer cluster points.

\[
F(W, Z) = \sum_{j=1}^{K} \sum_{i=1}^{N} w_{ij} d(x_i, z_j) \quad (1)
\]

\[
d(x_i, z_j) = \sum_{i=1}^{m} s_i(x_i, z_j) \quad (2)
\]

\[
w_{ij} \in \{0,1\}, 1 \leq j \leq K, 1 \leq i \leq n \quad (3)
\]

\[
\sum_{j=1}^{K} w_{ij} = 1, 1 \leq i \leq n \quad (4)
\]

\[
0 \leq \sum_{i=1}^{n} w_{ij} \leq k \quad (5)
\]

\[
d = \text{arccos}\left[\cos b_1 \times \cos b_2 \times \cos(a_i - a_j) + \sin b_1 \times \sin b_2\right] \times 6371 \quad (6)
\]

Equation (1) is the optimization function; equation (2) is used to calculate the similarity between sample \( x_i \) and cluster center \( z_j \); \( s_i(x_i, z_j) \) is the similarity between sample \( x_i \) and cluster center \( z_j \) on attribute \( s_i \); equation (3) is the membership relationship between sample \( x_i \) and class \( j \), if \( x_i \) belongs to class \( j \), \( w_{ij} = 1 \) Otherwise, it is equal to 0; equation (4) indicates that any sample point belongs to only one class; equation (5) indicates that there is at least one sample data in any class, but not more than all samples; if \( s_i \) is numerical data, then \( s_i(x_i, z_j) = \|x_i - z_j\| \); equation (6) indicates its Euclidean distance measure.

### 3.4. Clustering Validity Index

DBSCAN algorithm and K-means algorithm can only find spherical clusters with different characteristics. This algorithm can find clusters of any shape after eliminating noise points. The main factors influencing the clustering results are to select the set \( X \) of \( n \) C-end customer demand point data and the parameters \( \varepsilon \) and \( \text{Minpts} \).

### 3.5. DBSCAN Clustering Algorithm

For a clustering result, the closer the inner cluster is, the more scattered the clusters are, the better the clustering effect is.

1. Cluster internal distance
The inner distance of each cluster is the sum of the distances from all the demand points to the cluster center. The inner distance of clusters is obtained by adding the inner distance of $k$ clusters. The smaller it is, the closer the objects are. The formula is as follows:

$$D = \sum_{i=1}^{k} \sum_{x \in C_i} |x - z_i|$$

where: $D$ represents the internal distance of cluster, $x$ represents a demand point in cluster $C_i$, and $z_i$ represents the cluster center of cluster $C_i$.

(2) Average distance between clusters

The average distance between clusters is the average value of the sum of the distance from the cluster center of each cluster to the sample center of the whole demand point. The larger it is, the more scattered it is. The formula is as follows:

$$L = \frac{1}{k} \sum_{i=1}^{k} |z_i - z|$$

where: $L$ represents the average distance between clusters, $z_i$ represents the clustering center of the $i$-th cluster, $z$ represents the center of the whole set of demand points, and $k$ represents the number of clusters.

4. Empirical Analysis

In this paper, 125 demand points of Dongzhimen street in Dongcheng District of Beijing are randomly selected as experimental data (125 demand points are distributed in an area of 1 square kilometer), and the above two algorithms are used for clustering.

4.1. K-means Clustering Algorithm

For k-means algorithm, it is difficult to determine the number of clusters $k$. Here, it is assumed that $k$ is 10, 11, 19, by comparing the index values of clustering results, the optimal value of $k$ is determined. When $k$ takes different values, the statistical summary of parameters is shown in table 1.

From table 1, it can be seen that the internal distance of clusters decreases with the increase of $k$ value in general; the average distance between clusters fluctuates with the change of $k$ value without obvious downward or upward trend; the number of samples in the largest cluster decreases with the increase of $k$ value, and the difference between the number of samples in the largest and the smallest cluster decreases with the increase of $k$ value; the operation time increases with the increase of $k$ value.

To sum up, when the $k$ setting is small, the points in the cluster will not be compact enough, and the difference between the minimum and the maximum number of samples in the cluster will be large; when the $k$ setting is large, there will be only one sample point in some clusters, which is easy to produce outliers and affect the clustering effect. Under comprehensive consideration, the optimal number of classification to determine the demand point is 15.
Table 1. Output of K-means clustering result related parameters.

| k  | 10  | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Cluster internal distance (m) | 3 127 | 2 392 | 1 857 | 1 838 | 1 448 | 1 300 | 1 487 | 1 553 | 1 212 | 1 436 |
| Average distance between clusters (m) | 300 | 309 | 328 | 307 | 306 | 305 | 322 | 316 | 308 | 298 |
| Maximum number of samples in cluster | 22 | 25 | 21 | 16 | 21 | 21 | 15 | 17 | 15 |
| Minimum number of samples in cluster | 6 | 2 | 1 | 4 | 3 | 2 | 1 | 3 | 3 |
| Operation time (ms) | 18 | 17 | 18 | 20 | 21 | 21 | 22 | 24 | 22 | 23 |

4.2. DBSCAN Clustering Algorithm

Before the DBSCAN clustering algorithm clusters the demand points, it is necessary to determine the neighborhood radius \( \varepsilon \) and the density threshold \( \text{Minpts} \). As the sample data used in this paper is selected from Dongzhimen street, Dongcheng District, about 1 square kilometer, the sample data volume is 125. If the sample points are assumed to be evenly distributed in the map area, there is a demand point in a circular area with a radius of 50 meters. If the number of demand points in the region is greater than or equal to 2, the sample points in the circular region can be regarded as a cluster. Therefore, in this paper, \( \text{Minpts} = 2 \) is assumed to be the optimal value of \( \varepsilon \) by observing the change of clustering results, taking into account the flexibility of \( \varepsilon \), \( \varepsilon \) can be taken as 40, 50, 60, 70, 80 successively. When \( \varepsilon \) takes different values, the statistical summary of parameters is shown in table 2.

Table 2. Output of DBSCAN clustering result related parameters.

| \( \varepsilon \) | 40 | 50 | 60 | 70 |
|----------------|----|----|----|----|
| Cluster internal distance (m) | 2 030 | 4 572 | 7 227 | 12 194 |
| Average distance between clusters (m) | 285 | 288 | 318 | 343 |
| Maximum number of samples in cluster | 19 | 28 | 50 | 60 |
| Minimum number of samples in cluster | 2 | 2 | 2 | 2 |
| Noise number | 26 | 18 | 13 | 10 |
| Operation time (ms) | 16 | 15 | 15 | 15 |

As can be seen from table 2, when \( \text{Minpts} \) is constant, with the increase of \( \varepsilon \), only the number of noise points decreased, and other values increased. Therefore, when \( \varepsilon \) is larger, some clusters contain too many samples, and the distribution of sample points in the cluster is not close, so it is not suitable to be a customer cluster. On the contrary, when \( \varepsilon \) takes a smaller value, the number of sample points in the cluster will not be too many, and the internal objects of the small cluster composed of these points are very close, but at the same time, the algorithm will also classify some points that are not close enough as noise points, resulting in more noise points. Under the comprehensive comparison, the clustering results under \( \varepsilon = 50 \) and \( \text{Minpts} = 2 \) can be selected as the distribution results of customer cluster points.
4.3. Comparison of Clustering Results

In order to compare the clustering effect of the two algorithms, two groups of results with the same number of classifications and the best clustering effect of the algorithm are selected for comparison.

In DBSCAN algorithm, when $\varepsilon = 50$, $Minpts = 2$, the number of customer points is 17; in k-means algorithm, when $k = 17$, the number of customer points is 17. The comparison between the two cases is shown in figure 1.

![Comparison of clustering results between DBSCAN algorithm and K-means algorithm](image1.png)

**Figure 1.** Comparison of clustering results between DBSCAN algorithm and K-means algorithm.

It can be seen from figure 1 that when the number of clusters is 17, the internal distance of K-means clustering is reduced by 3019 meters compared with DBSCAN; the average distance between K-means clustering is increased by 61 meters compared with DBSCAN; the difference between the maximum and minimum number of samples in K-means clustering is 12, and DBSCAN algorithm is 26. To sum up, compared with DBSCAN clustering, K-means clustering has smaller internal distance and closer sample points in the cluster; larger average distance between clusters and more scattered between clusters; smaller difference between the number of samples in the largest and smallest clusters, and more uniform distribution of cluster size. Therefore, K-means clustering is more effective when the number of classification is the same.

Through the analysis of the clustering results of the two algorithms, we can see that in k-means algorithm, the best clustering effect is $k = 15$; in DBSCAN algorithm, the best clustering effect is $\varepsilon = 50$ and $Minpts = 2$. The comparison between the two is shown in figure 2.

![Comparison of clustering results between DBSCAN algorithm and K-means algorithm](image2.png)

**Figure 2.** Comparison of clustering results between DBSCAN algorithm and K-means algorithm.
From figure 2, it can be seen that the internal distance of K-means clustering is 3,272 meters less than that of DBSCAN; the average distance between K-means clustering is 17 meters more than that of DBSCAN; the difference between the maximum and minimum number of samples in K-means clustering is 19, and DBSCAN is 26. To sum up, K-means algorithm is still better even if the results of the two clustering effects are the best.

4.4. Construction Ideas of Terminal Service Outlets
Before In the urban distribution network, the terminal service outlets are directly facing the customers, which often need to be arranged in the place where the customers are concentrated. In table 3, the proportion of clusters with 1-5, 11-15 and 20-25 sample points is 33%, 20% and 7% respectively, so the number of clusters with different sample points is also quite different. In addition, the number of sample points in the largest cluster is 21, and the number of sample points in the smallest cluster is 2. The difference between them is 19. It can be seen that the number of sample points in clusters of different sizes also has a large gap. To sum up, the demand points served by the end service outlets are unevenly distributed. There are some regional demand points with dense distribution and large demand, and some regional demand points with scattered distribution and small demand. Therefore, considering the limited service radius of the end-to-end service nodes, the uneven distribution of demand points will lead to a large difference in demand within the service areas of different nodes. By clustering the C-end customers, we can analyze the “social” characteristics of the demand, and then carry out zoning planning.

| Number of sample points in the cluster | Number of clusters | Proportion |
|----------------------------------------|-------------------|------------|
| 1-5                                    | 5                 | 33%        |
| 6-10                                   | 6                 | 40%        |
| 11-15                                  | 3                 | 20%        |
| 16-20                                  | 0                 | 0%         |
| 20-25                                  | 1                 | 7%         |

5. Conclusion
In this paper, it is difficult to determine the end distribution demand point, K-means and DBSCAN are used to cluster the demand points, and two indexes are selected to select the best. Through the empirical analysis, the clustering methods are compared, and the clustering effect of C-terminal demand points is discussed, which provides reference for terminal service network location. In the future work, we will further discuss the clustering effect of dynamic grid density algorithm and hierarchical clustering algorithm on C-end demand points.

Acknowledgement
The research is supported by grants from National Key R&D Program of China (2018YFB1601600).

References
[1] Li M, Zhang J and Zhou X P 2014 Construction of community logistics distribution network based on K-means algorithm Applied Mechanics & Materials 641&642 1271-1274
[2] Lan Y H and Li Y S 2016 The method of cluster and entropy apply in city distribution center International Conference on Educational and Network Technology
[3] Zhu X Z 2016 Research and Application of Hot Spots in Scenic Spots Based on Spatial Data Mining (Wuhan: Central China Normal University)
[4] Wen H Y, Lu D Y and Wu Y P 2018 Division of logistics park traffic zones based on the improved fuzzy clustering method Journal of Harbin Institute of Technology 50 (3) 103-108
[5] Bychko V A, Yershov R D and Plichko L 2018 Automation of clustering methods at planning logistics processes with georeferenced *IEEE 9th International Conference on Dependable Systems, Services and Technologies* DOI: 10.1109/DESSERT.2018.8409165

[6] Zhou X, Xu M Z and Lv Q G 2018 A model of gradual coverage and site selection of self-raised points based on the distribution of customer points *Computer Integrated Manufacturing System* 24 (11) 231-240

[7] Wang P F 2016 Research on the Distribution of Express Service Outlets Based on Clustering Algorithm (Chengdu: Chengdu University of Technology)

[8] Nam N M, Rector R B and Giles D 2015 Minimizing differences of fonvex functions with applications to facility location and clustering *Journal of Optimization Theory & Applications* 1-24