Application of Set Covering Model: Site Selection of Express Wrappage Recycling Centers in Central Chengdu

Liu Haikang

International Economics and Trade (Bilingual) 41802010

A R T I C L E I N F O

Keywords:
Reverse logistics
Set covering model
Cluster analysis
Greedy algorithm

A B S T R A C T

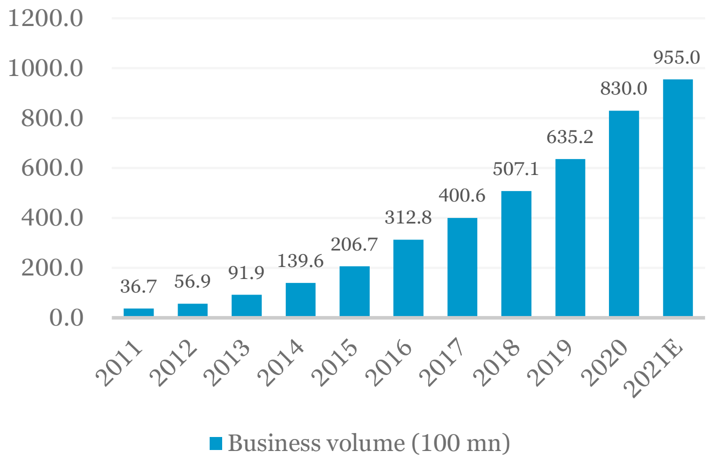
In recent years, the rapid increase in the express businesses in China has led to huge amount of recyclable wrappages being discarded and a severe waste of resources, though it is an embodiment of a pull in consumer demand. A number of researchers have deepened their study on reverse logistics with the view of solving the problem. In terms of the siting of recycling points in express wrappage recycling network, this essay formulates a set covering model and applied in the location selection practice of central Chengdu's express packaging recycling points, then solved the model with K-means clustering algorithm and greedy algorithm, eventually obtained the optimal solution with all express outlets covered by the fewest recycling points, which would be of some practical value of reference.

Content

1. Introduction

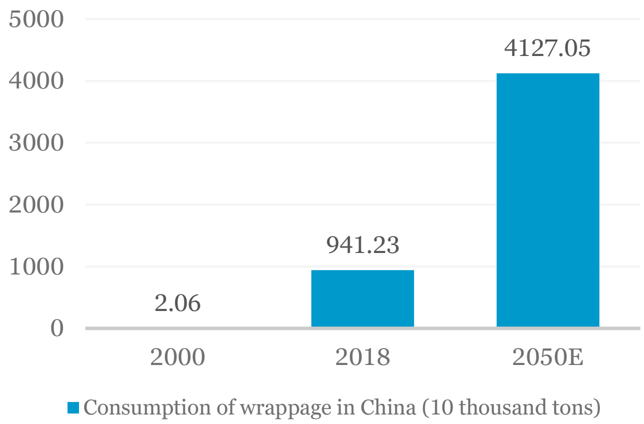
In 2020, the express delivery volume in China has exceeded 80 billion, 30.7% higher in comparison to 2019[1]. As is shown in Figure 1.1, the business has been burgeoning since 2011[2].

Figure 1.1 Business volume of express services in China (2011-2021E)



According to the statistical indicators for express service in the *Thirteenth Five-Year Plan*, the consumption of express packaging materials, also known as wrappage, has increased from 0.01 in 2000 to about 50 in 2020, per resident annually. In China, suggested by the international NGO *Greenpeace*, the consumption of various express packaging materials has soared from 20.6 thousand tons to 9.41 million tons during 2000-2018. Provided that the express volume continues to grow at this rapid pace, the very figure is predicted to be about 41.27 million tons in 2050 in China (shown in Figure1.2), which will pose extreme threat to the environment and resources, unless measures are enforced to control the usage of wrappages[3].

Figure 1.2 Consumption of wrappage and prediction in China



In recent years, with a view to solve the problem, plentiful scholars have conducted a lot of research in the field of "reverse logistics". According to the *National Standard Logistical Terminology (GB/T 18354-2006)* released by Ministry of Transport of China, reverse logistics refers to the logistics activities triggered by the movement of items from downstream to upstream in the supply chain[4].

In terms of the construction of the reverse logistics network, *Ma Jianlong et al.*[5] built a dynamic location model for solid waste reverse logistics based on a multi-period and multi-objective perspective with the goal of minimizing the total cost, and studied how to determine the location of solid waste recycling point, transfer point and treatment station in each period, as well as the allocation of waste generation point. *Mu Jingjing*[6] took express packaging recycling in 16 municipal districts of Shanghai as the research objects, and used cluster analysis to analyze the consumption level and residents' purchasing power in each region, based on which the express packaging recycling center was classified and the construction scale was determined. *Zhou Xiaoye et al.*[7] used an improved K-means clustering method combined with a set covering model to study the location of the recycling point, with the least recycling points to cover all express outlets, which reduces the cost of network recycling and improves the efficiency of express wrappage recycling.

The research hitherto on the recycling of express packaging at home and abroad mainly focuses on the recycling process and mode, and there are relatively few studies on the location of the express wrappage recycling point. Enlightened by the work of *Zhou Xiaoye*, this paper draws on the model and improves it, and proposes a K-means clustering set coverage location model based on the greedy algorithm, and uses 5 central districts in Chengdu (Jinjiang, Qingyang, Wuhou, Jinniu and Chenghua) to conduct site selection experiments. The results prove that this model can more effectively save the construction cost of recycling points, and further improve the efficiency of express wrappage recycling.

2. Model Design

2.1. Methodology

The main function of the recycling centers is express package collection and short-term storage. Existing express packaging recycling is done by express companies in their own way. Simple recycling is carried out through the express outlets of their respective companies, and only wrappage that can be directly reused is recycled, resulting in high recycling costs. The recycling points at issue in this paper are made responsible for collecting the wrappage remained in all express outlets (defined as demand points) within a certain range of recycling radius. Each express company can use the recycling point for wrappage recycling, realize the sharing of recycling points and make full use of the existing logistics facilities, so that the recycling cost for the entire society can be reduced.

2.2. Assumptions

i. Consider establishing cost only, i.e., other costs and expenses such as operating cost are ignored.
ii. The cost to set up each point is equally fixed.
iii. Establishing cost of one recycling point is unrelated to the number of express outlets it serves.

2.3. Set Covering Model (SCM)

The set covering problem is a typical one of discrete mathematics and is one of the NP-complete problems put forward by *Richard Karp*. The SCM has been frequently applied in the site selection practice of service outlets, logistic distribution center and delivery station. The underlying idea is to formulate a scientific and reasonable equipment placement plan to meet the needs of each demand point under the premise that the demand of each demand point is known to be limited. Therefore, the scheme engenders the minimum facility construction cost and each demand point is served by at least one facility[8]. Mathematically, the model can be expressed in linear-programming form as Equation set2.1.

Equation set2.1

$$\min Z = \sum_{j=1}^{M} f_j x_j \qquad (2.1a)$$

$$s.t.\begin{cases} \sum_{j=1}^{M} c_{ij} x_j \geqslant 1,\ i = 1, 2, \ldots, N & (2.1b) \\ x_j \cdot dis(d_i, s_j) \leqslant R & (2.1c) \\ c_{ij} = \begin{cases} 1,\ demand\ point[i]\ is\ covered\ by\ facility[j] \\ 0,\ otherwise \end{cases} & (2.1d) \\ x_j = \begin{cases} 1,\ construct\ facility\ at\ candidate\ point[j] \\ 0,\ otherwise \end{cases} & (2.1e) \end{cases}$$

where f represents the fixed cost to establish a facility at point j, x is a binary control variable; equation 2.1a is the objective function, 2.1b means each demand point is covered by at least one facility, 2.1c limits the distance between the very facility, *s*, and the demand point, *d*, to *R*, which is given. There are *M* candidate points, the locations at which whether facilities will be establish is to be discussed, and *N* demand points.

Normally, a set covering problem is solved by means of linear programming, cluster analysis and greedy algorithm. For instance, the author of reference [8] used linear programming to solve the set covering problem with MATLAB.

In this paper, the main methods involved are K-means cluster analysis and greedy algorithm.

2.4. K-Means Cluster Analysis (KMCA)

K-means cluster analysis is a dynamic clustering method proposed by *James Macqueen* in 1967. It supposes that all individuals in the sample can be divided into C categories, and C initial cluster centers are selected. Then, based on the minimum distance principle, each individual can be assigned to a certain category. Next, iteratively calculate the cluster centers of each category and adjust the clustering situation according to the new cluster centers until the iteration converges or the cluster centers no longer change[9]. This way, each cluster center will be used as a retained candidate point, and due to the instability of this multivariate statistical algorithm, the position of the cluster center calculated each time is not the same, so the position of the initial candidate point is also not certain.

In our experiment, the determination of the optimal number of clusters is of crucial importance. Too few clusters will result in incomplete coverage, and too many clusters will result in excessive construction costs and waste of resources. Therefore, I will first analyze the relationship between the maximum distance between the cluster center and each demand point in the cluster (negatively correlated), which is determined by the number of different clusters; then, given a threshold R, with the maximum distance within the cluster does not exceed R as the goal, I would determine the minimum number of clusters that can cover all demand points, i.e., the optimal number of clusters. After determining the optimal number of clusters, I would perform cluster analysis on the sample, and extract the coordinates of each cluster center, which is used as the initial cluster center (initial location point).

2.5. Greedy Algorithm (GA)

Greedy algorithm is a scheme which always makes the best choice in the current view when solving a problem. That is, without considering the overall optimality, the algorithm obtains a local optimal solution in a certain sense[10].

After using KMCA to determine the initial cluster centers, I've found that the number of these cluster centers is more than necessary and the distribution density are excessively high, and there is a serious problem of repeated coverage. If these initial points are used as the basis for site selection, it will lead to inefficient use of resources and drive up construction costs. To solve this problem, I would improve the model based on the GA algorithm idea. The logic chain is as follows:

i. For each initial cluster center in the centers set C, find its covering set given a radius R. Calculate the number of demand points it covers and record the times for which the points involved are covered as array T.

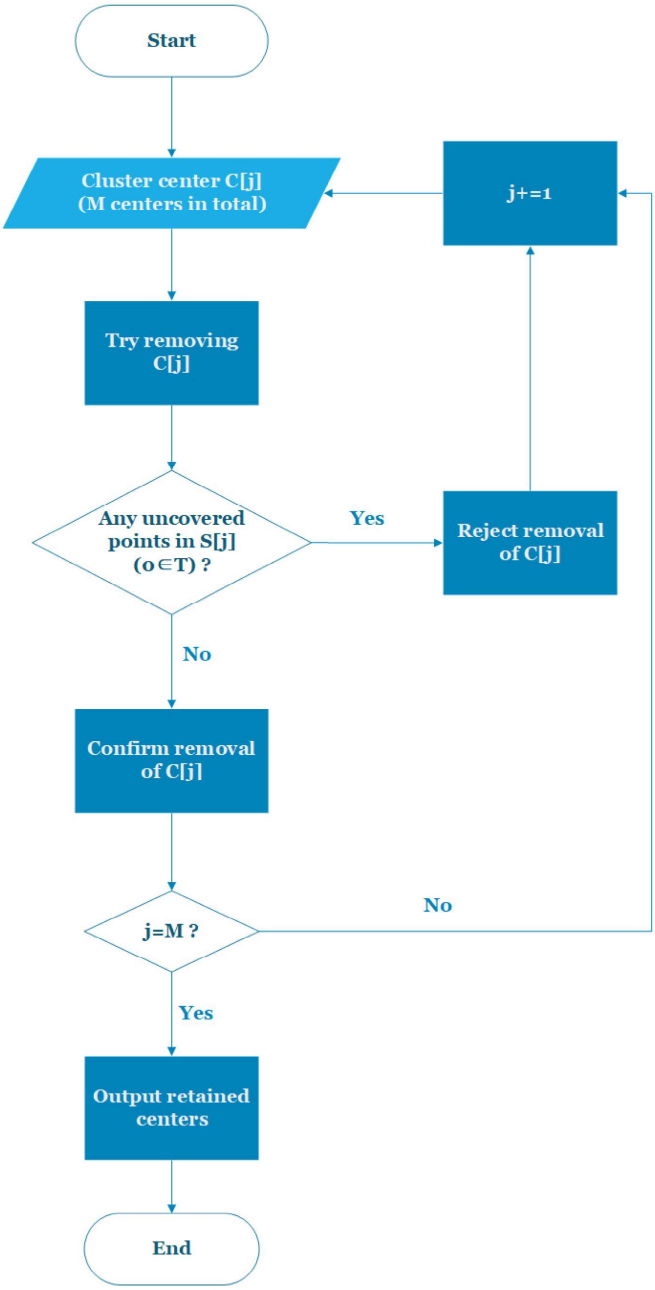
ii. Rank all initial centers, in an ascending order of the quantity each center covers. Designate rC to represent the ranked initial centers set, the first in which C_1 is the center with least covering quantity.

iii. Starting from C_1, one by one, try removing initial centers, whose served demand points are included as set S_j. If C_1 was removed, the times for which the demand points in S_1 are covered, accordingly reflected as numbers in array T, would be reduced by 1. Then judge if 1) No element in T appears to be 0, confirm the removal; 2) One or more elements in T are reduced to 0, reject the removal and return to the pre-removal stage, and skip this center.

iv. Keep doing so until all initial centers have been tested. In the end, the remaining ones are necessary cluster centers which cannot be removed. These centers satisfy two requirements: 1) Each of the demand points is covered by at least one center; 2) Clusters are kept as few as possible. That is to say, the quantity of points that each cluster center covers are as many as possible.

Such a procedure can be shown by a flow chart in Figure2.1. In short, the embodiment of the idea of greedy algorithm in the model is that those cluster centers with little coverage will be deleted first. Consequently, cluster centers remained are those with more coverage.

Figure 2.1 GA procedure



3. Practical Application

3.1. Experiment Profile

- Sample selected: All express service outlets in 5 central districts of Chengdu (Jinjiang, Qingyang, Wuhou, Jinniu and Chenghua)
- Sample capacity: 3,438
- Data source: Amap
- Platform: Python3.8 (Jupyter)

3.2. Data Scraping

POI (point of interest) refers to the point data in online electronic map, which basically contains four attributes: name, address, coordinates and category of a location spot. Based on the POI code of Amap (updated on August 10th, 2017), I designate the three types of data shown in Table3.1 as the category of express outlets. Then I scraped the needed data of 5 central districts of Chengdu using Amap's WEB service API, and the result sample has been stored as a spreadsheet in Appendix1.

Table 3.1 Amap POI code defined as express outlets

ID	NEW_TYPE	Big Category	Mid Category	Sub Category
387	070401	Daily Life Service	Post Office	Express Post
388	070500	Daily Life Service	Logistics Service	Logistics Service
389	070501	Daily Life Service	Logistics Service	Logistics Warehouse Space

3.3. Visualization

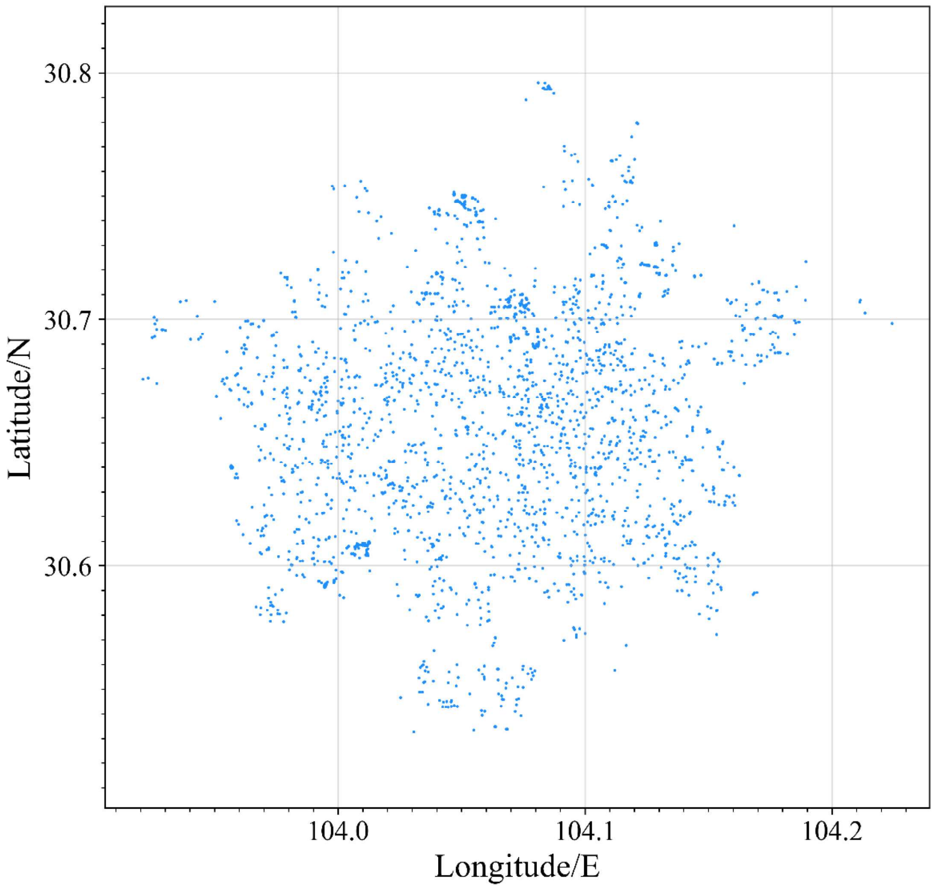
3.3.1. Visualize with online map

The sample data can be visualized on the map. Here I deploy the Folium module to achieve this. The map has been stored as a HTML webpage in Appendix2.

3.3.2. Visualize with Cartesian coordinates

To keep things simple, a Cartesian coordinate system is more suitable to illustrate the data. With Matplotlib module, the express outlets are abstracted as points in the coordinate graph in Figure3.1, with the ordinate being latitude and the abscissa being longitude.

Figure 3.1 Distribution of express outlets



3.4. KMCA

3.4.1. Choosing optimal quantity of clusters

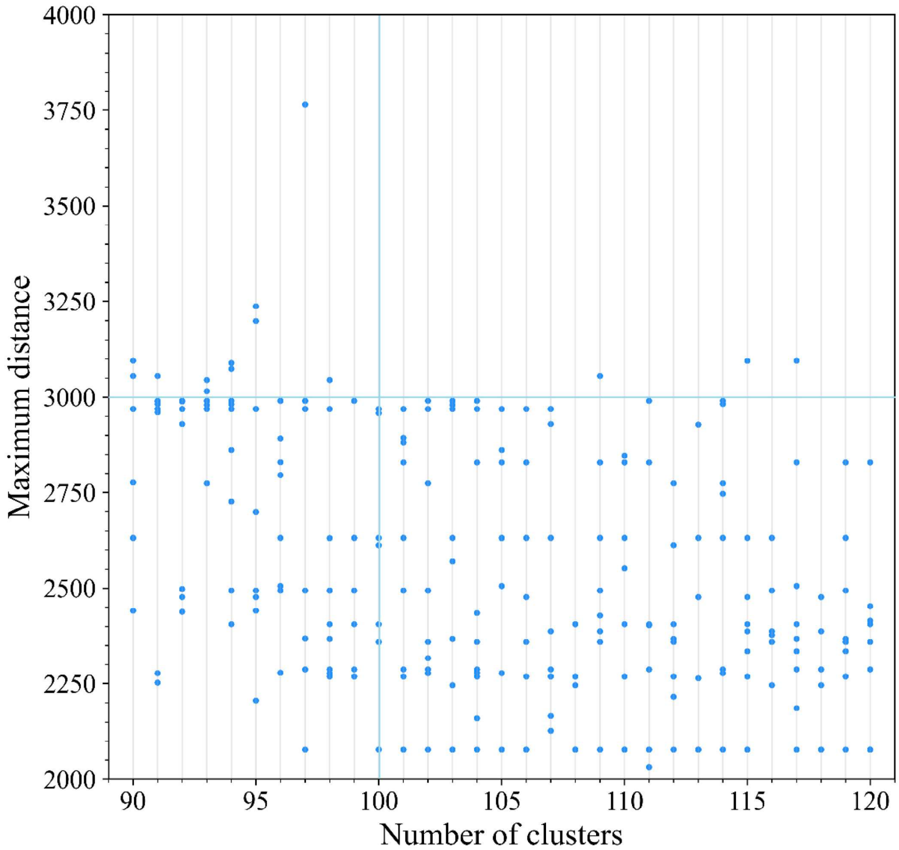
Let the number of clusters be n and the maximum distance within a cluster (the distance between the center and the farthest point) be $Maxdis_n$. According to the *Technical Code of the Transfer Stations of Household Waste (CJJ/T47-2016)* released by the MOHURD of China, the service radius of a transfer station should be 3~5 kilometers[11]. Thus, 3 km is chosen as the coverage radius R of the recovery point, which is converted into a geographical coordinate distance of about 0.027° using the conversion formula specified in Equation3.1.

Equation 3.1

$$D = R \cdot \arccos\left[\cos y_1 \cos y_2 \cos(x_1 - x_2) + \sin y_1 \sin y_2\right]$$

Since the number of sample points is enormous and the distribution is so much scattered, the obvious clustering cannot be seen from the figure. In order to reduce the number of iterations, I select n=90 as the initial number of clusters and give step =1 for iteration. The goal is to keep $Maxdis_n$ lower than R. Due to the instability of KMCA, I've repeated the calculation of the maximum distance 10 times for each n, and analyzed the probability of falling point. The result is shown in Figure3.2, from which it is safe to say that most probably, when the number of clusters is greater than 100, the maximum distance within a cluster can be effectively controlled within 3 km, so n=100 is selected as the optimal number of clusters.

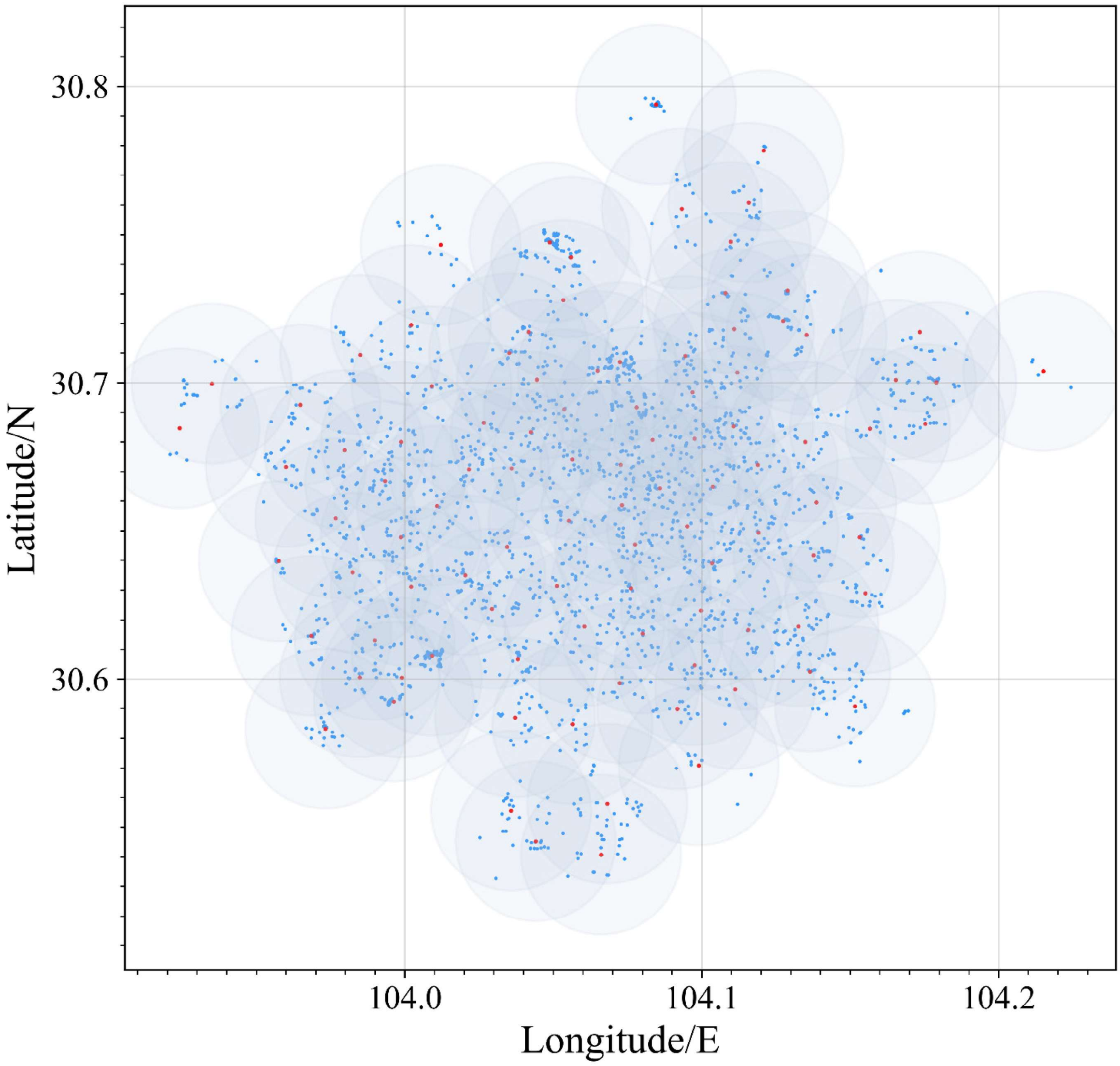
Figure 3.2 Maximum distance & Cluster quantity correlation



3.4.2. Cluster analysis

Scikit-learn is deployed here to conduct the cluster analysis to the sample. The initial covering result is shown in Figure3.3, where each red point represents a cluster center and the circle with a radius of 0.027 is the coverage of the center.

Figure 3.3 Initial covering



We can see that the KMCA designates 100 cluster centers and it is true that all express outlets are covered. However, the limitation is also obvious, i.e., excessively high covering rate, low resource utilization and high construction cost. It can be calculated that a cluster covers an area of 28.27 km^2, hence 2,827 km^2 for 100 of them. Whereas the area of the 5 districts totals to 423.86 km^2, the utilization rate of the recycling points designated by the candidate centers is only 15% (although we cannot simply use the area covered as a measurement of utilization). Obviously, it is necessary to screen out some of the candidate points, thereby reducing the number of candidate points, repeated coverage and construction costs and improving utilization.

3.5. Improvement by GA

3.5.1. Model description

The use of GA algorithm to improve the SCM mainly focuses on two goals:

i. The reserved cluster centers can cover all sample points;

ii. Keep the number of cluster centers as small as possible (keep as much as possible those cluster centers that cover as many points as possible).

Therefore, the GA algorithm model can be expressed in mathematically as Equation set3.2.

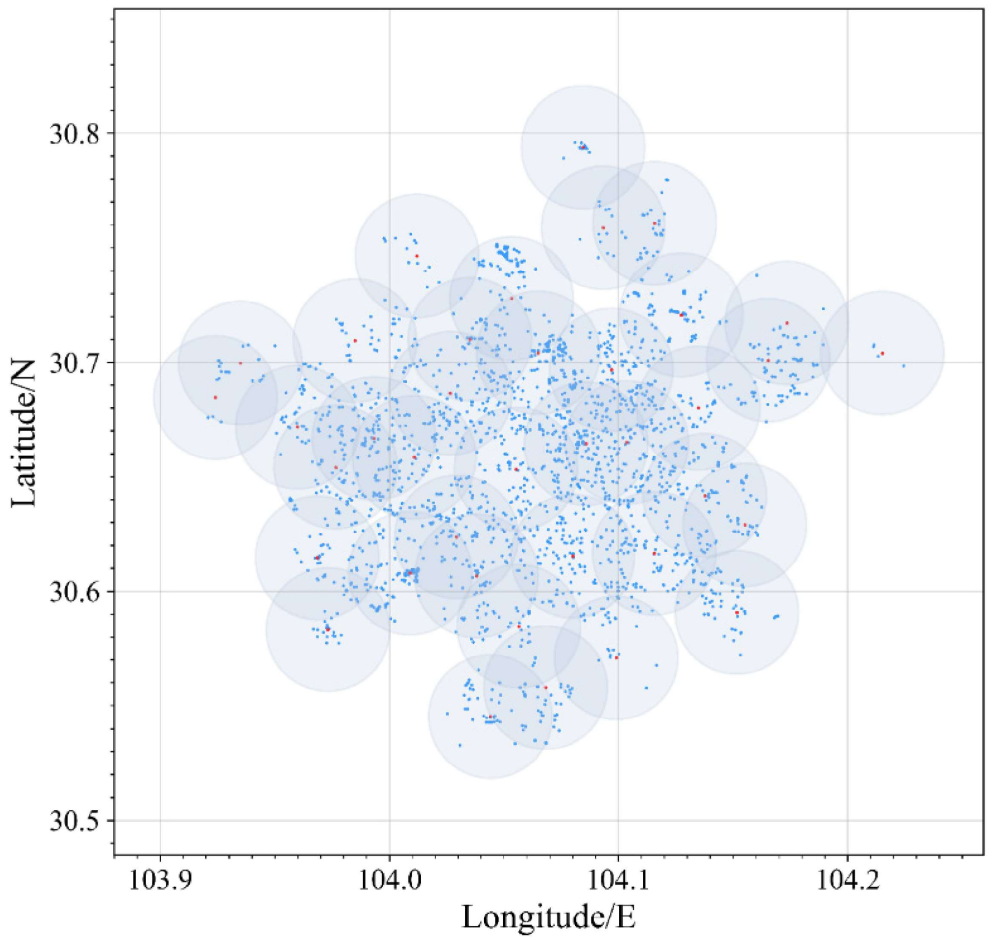
Equation set3.2

$$\begin{cases} \min \sum_{j=1}^{100} x_j & (3.2a) \\ \max \sum_{i=1}^{3438} N_{ji},\ j=1,2,\dots,100 & (3.2b) \end{cases}$$

$$s.t. \begin{cases} \sum_{j=1}^{100} c_{ij} x_j \geqslant 1,\ i=1,2,\dots,3438 & (3.2c) \\ x_j = \begin{cases} 1,\ retain\ cluster\ center[j] \\ 0,\ otherwise \end{cases} & (3.2d) \\ c_{ij} = \begin{cases} 1,\ dis(d_i,s_j) \leqslant R \\ 0,\ otherwise \end{cases} & (3.2e) \\ N_{ji} = \begin{cases} 1,\ c_{ij}=1 \\ 0,\ otherwise \end{cases} & (3.2f) \end{cases}$$

where d_i, s_j represent demand points and cluster centers respectively ($i=1,2,\dots,3438$, $j=1,2,\dots,100$). $dis(d_i,s_j)$ is the distance between d_i and s_j. x_j and c_{ij} are binary variables, the latter of whom is a judgment over whether d_i is covered by s_j. R equals 3 km or 0.027° . $\sum N_{ji}$ counts the quantity of points s_j covers. Equations 3.2a and 3.2b are objective functions, and they are equivalent in a sense. 3.2c defines the constraint that guarantees each demand point is covered by at least one cluster center.

Figure 3.4 Improved covering



3.5.2. Result analysis

After improvement, the covering result is shown in Figure3.4 and visualized in another HTML webpage of map in Appendix2, in which only 38 clusters are retained*, represented by circles with red centers in the graph. Evidently, only 38 recycling points are necessary to serve all express outlets, and the utilization rate also increases by about 167%. The detailed information of these centers has been saved in a spreadsheet that can be found in Appendix1.

4. Limitations

The KMCA & GA based SCM model has performed quite well in screening out unnecessary cluster centers. However, the algorithm also has some limitations.

4.1. Problems with Premises

The model has three assumptions in practical application:

i. Consider establishing cost only, i.e., other costs and expenses such as operating cost are ignored.
ii. The cost to set up each point is equally fixed.
iii. Establishing cost of one recycling point is unrelated to the number of express outlets it serves.

Obviously, these assumptions are also the crux of the model.

For assumption 1), in reality, if the recycling point is not a public facility constructed by the government but by private firms, then the company must consider the benefits that it can obtain when placing the recycling point. ROI is an important indicator for companies to measure the value of an investment project. Talking about input and ignoring output is clearly not in the interest of the company. For example, for a recycling center with only a small number of express points covered, the company needs to assess whether it is worthwhile to set up a packaging collection point exclusively for these delivery outlets, or just include them in the service scope of other nearby recycling points.

As with assumption 2), the construction costs of different recycling points are different. One thing is that the land prices and rents of different locations are different, resulting in diversified fixed costs; the other is that the express delivery industry, as a service-oriented industry, has a large proportion of variable costs in its cost structure, which leads to the establishing expense and operation overhead varying greatly from one recycling center to another.

In terms of hypothesis 3), theoretically, the set-up cost of a recycling center is positively related to its service volume, that is, the more express outlets are within the service scope of a recycling point, the larger the construction scale and the higher the establishing cost. This will make the problem more complicated: if the construction cost is related to the service volume of the recycling points, then it is not that the fewer recycling points the better, because reducing the number of recycling points will increase the service volume of the recycling points, so it is necessary to balance the two factors in order to achieve cost minimization.

4.2. Problem of Distortion

The SCM uses abstract coordinates data, and the distance between paired points is not the real distance. There are two reasons: first, the model uses the straight-line distance between points, and the coverage radius is not the actual distance. That is to say, there may be such a situation: although the model calculates that the express point is within 3 km of the recycling center, the geographical distance may not be 3 km, which causes a certain degree of deviation of the model from the real situation; second, there will inevitably be errors when the latitude and longitude coordinates are converted to the spherical distance, which leads to the distortion of the distance.

* *Due to instability, the result may vary.*

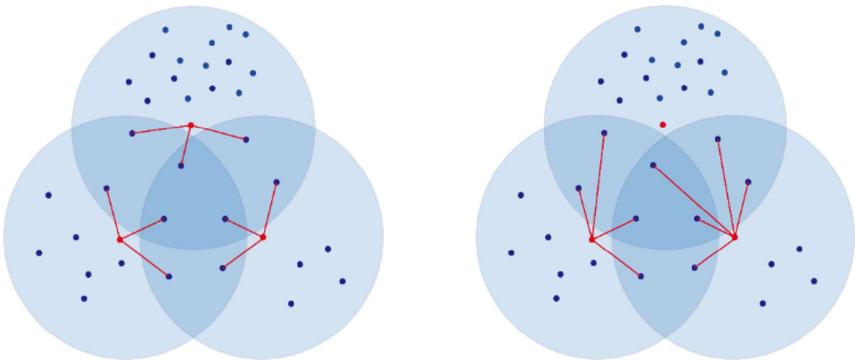
4.3. Problem with Overly Covered Outlets

Even if the SCM is used to improve the model that the number of collection points can be greatly reduced, there are still express outlets that are covered more than once, namely overly covered outlets. Which recycling center should these points be allocated to? I've come up with two solutions:

i. Assign according to distance: for each overly covered point, assign the closest recycling point to provide the collection service of express wrappage for the point, as shown in Figure 3.5a;

ii. Assign according to workload: that is, for each overly covered point, evaluate the service volume (workload) of the surrounding recycling points (that is, the number of express outlets in the coverage area except for the overly covered points), find the recycling centers with least workload and then assign the overly covered points to such centers based on the principle of the shortest distance, as shown in Figure3.5b.

Figure3.5 Problems with overly covered points



(a) (b)

In comparison, the first option is relatively simple, but it is easy to cause too much workload to some recycling points; the second one is more reasonable, but the algorithm is more complicated.

In summary, the model designed in this article only takes into account the cost factor and assumes that the cost of each point is the same. In fact, the private suppliers of these recycling points will also measure the investment return, and the construction costs of different locations are also different. Besides, the straight-line distance is also not practically since when considering the actual distance, it is also necessary to consider factors such as road traffic conditions. After all, the model can offer some reference for site selection in real life, but it needs further improvement to be applied more realistically and more flexibly.

5. Conclusion

Aiming at the social problem of the recycling of express packaging materials, this paper takes 5 central districts in Chengdu as an example to study the site selection of recycling centers from a macro perspective, with set coverage model, K-means cluster analysis and improved greedy algorithm. The model has been implemented through a Python program mainly designed and coded by myself, with a few modified after referred to related materials. Practice has proven that the site selection model designed in the article has certain reference value, but it is still a relatively preliminary one. As a matter of fact, more realistic considerations need to be incorporated in the application to make it adaptable to various changes, more universal and more practical.

Reverse logistics of solid waste such as express packaging materials is an important research field for building a resource-saving and environment-friendly society, and it concerns the vital interests of every social participant. It is not enough to rely on the government and enterprise-level management planning for the recycling and utilization of domestic waste. At the same time, it is also necessary to raise the awareness of all citizens to promote the reduction of domestic waste at the source and realize the participation of, and benefits to, all people.

Appendices

Appendix1: Spreadsheets
Appendix2: HTML webpages
Appendix3: Python program

Please see the folders attached to this paper.

References

[1] 中国政府网. 我国快递年业务量首次突破 800 亿件[EB/OL]. http://www.gov.cn/xinwen/2020-12/21/content_5572045.htm, 2020-12-21/2021-6-10.

[2] 中华人民共和国国家邮政局. 2019 年中国快递发展指数报告[DB/OL]. http://www.spb.gov.cn/xw/dtxx_15079/202003/t20200327_2068989.html, 2020-03-27/2021-6-10.

[3] 前瞻产业研究院. 2020 年中国快递行业市场现状及发展前景分析; 未来包装材料消耗量将突破 4000 万吨[EB/OL]. https://bg.qianzhan.com/trends/detail/506/200401-0e108353.html, 2020-04-01/2021-6-10.

[4] 全国物流标准化技术委员会. 国家标准《物流术语》GB/T 18354-2006 [DB/OL]. http://zizhan.mot.gov.cn/zhuantizhuanlan/gonglujiaotong/shoufeigongluzmk/zhengcefagui/201508/t20150814_1863913.html, 2015-08-14/2021-6-10.

[5] Ma Jianlong, & Jiang Jingqiu (2020). Research on Reverse Logistics of Municipal Solid Waste and Saving Environment Governance Cost [J]. *Price: Theory & Practice, 7, pp. 77-80.*

[6] Mu Jingjing (2020). Research on Classification of Express Packaging Recycling Center based on Cluster Analysis [J]. *Logistics Engineering and Management, 42(11), pp. 131-133.*

[7] Zhou Xiaoye, Ren Guibin, Ma Xiaoyun, Zhu Meilin, & Zhong Yu (2020). Location Selection of Recycling Points in Express Packaging Recycling Network [J]. *Packaging Engineering, 41(11), pp. 201-208.*

[8] Tan Zhuge (2016). Research on Express Cabinet Location Problem [D]. *Hefei, Anhui: Hefei University of Technology.*

[9] Si Shoukui, & Sun Xijing (2020). Python Mathematical Experiment and Modeling [M]. *Beijing: Science Press.*

[10] Kang Guihua (2016). Introduction to Computation [M]. *Beijing: China Railway Publishing House.*

[11] 中华人民共和国住房和城乡建设部. 住房城乡建设部关于发布行业标准《生活垃圾转运站技术规范》的公告 [EB/OL]. http://mohurd.gov.cn/wjfb/201607/t20160715_228156.html, 2016-06-14/2021-6-10.