Modeling and Simulation 建模与仿真, 2023, 12(3), 1899-1907
Published Online May 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.123174

Hans汉斯

基于K-Means + MST算法的疫情期间生活物资科学管理预测

——以长春市朝阳区物资发放点为例

朱　寅，张　嵘，李　帅

上海理工大学机械工程学院，上海

收稿日期：2023年2月20日；录用日期：2023年5月3日；发布日期：2023年5月10日

摘　要

在疫情发生的封闭管理期间，针对不同的人群需要不同的管理方法，考虑到人口的规模及地理位置的差异，需要统筹各方因素建立一个行之有效的管理模式，主要考虑的典型问题有：生活物资的发放问题；物资运送路线问题。生活物资发放问题主要考虑物资投放点数量及位置的合理性，物资发放时人员交流接触尽可能少，减少病毒交流扩散的可能性，物资运送路线问题主要考虑没有一种特定合理的运输路线。本文使用高斯插值法对没有统一发放物资时期的感染人数进行拟合，并通过拟合好的函数对后面的感染人数进行预测，针对疫情期间运输路线不合理，没有合理的运输路线问题，通过K-means算法求出合理的物资投放点坐标并利用求出的坐标与MST算法相结合，得出较为合理的配送路线图，可作为以后发生灾害的参考路线以实现经济效益的最大化。

关键词

科学管理，高斯拟合，聚类，K-Means算法，MST算法

Scientific Management Prediction of Living Materials during the Epidemic Based on K-Means Algorithm and MST Algorithm

—Taking the Example of the Material Distribution Point in Chaoyang District, Changchun City

Yin Zhu, Rong Zhang, Shuai Li

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 20th, 2023; accepted: May 3rd, 2023; published: May 10th, 2023

文章引用: 朱寅, 张嵘, 李帅. 基于 K-Means + MST 算法的疫情期间生活物资科学管理预测[J]. 建模与仿真, 2023, 12(3): 1899-1907. DOI: 10.12677/mos.2023.123174

Abstract

During the closed management of the epidemic, different management methods are needed for different groups of people. Considering the size of the population and the differences in geographic location, it is necessary to establish an effective management model by integrating all factors. The main issues to be considered are: the issue of material distribution; the issue of material transportation routes. The main consideration of the distribution of living materials is the reasonableness of the number and location of the material delivery points, and the possibility of virus exchange and spread is reduced by minimizing the personnel exchange contact during the distribution of materials, the problem of material delivery routes mainly considers the absence of a specific and reasonable transportation route. In this paper, we use Gaussian interpolation method to fit the number of infected people in the period of no uniform distribution of supplies and predict the number of infected people later through the fitted function. For the problem of unreasonable transportation routes and no reasonable transportation routes during the epidemic, we use K-means algorithm to find out the reasonable coordinates of the material delivery points and use the found coordinates combined with MST algorithm to come up with a more reasonable distribution. The route map can be used as a reference route for future disasters to maximize the economic benefits.

Keywords

Scientific Management, Gauss Interpolation, Clustering, K-Means Algorithm, MST Algorithm

1. 引言

全国出现多次疫情爆发事件，本文以长春市为例，疫情期间的蔬菜物资发放成为焦点问题，发放方式不当很有可能造成二次传播。为了利于以后的疫情防控工作，本文使用高斯插值和 K-means 算法对疫情期间物资的科学管理方式对疫情的影响进行探索，实现了对生活物资投放点数量，位置的优化，可以为以后大规模封控情况下居民物资有序发放提供参考。

2. 模型的建立与求解

本文用到的符号及含义如表 1 所示：

Table 1. Symbols and meanings
表 1. 符号及含义

符号	含义
Q_i	在一天内 i 地区所需要的蔬菜包重量
D_i	第 i 个区域的小区坐标集合
X_i	第 i 个区域的小区 X 坐标集合
Y_i	第 i 个区域的小区 Y 坐标集合
X_i^{T}	第 i 个区域的小区 X 坐标集合的转置
α	表示在一天内 i 地区每个群众对应消耗生活物资的数量

Continued

λ_i	表示在 i 地区需要蔬菜的人口总数(不包含感染人数)
Y_i^{T}	第 i 个区域的小区 Y 坐标集合的转置
C	划分的簇
d_{ij}	i 投放点到 j 投放点的距离

2.1. 高斯拟合

2.1.1. 高斯拟合概述

高斯拟合是使用形如

$$Gi(x) = Ai * \exp\left((x - Bi)\text{^}2/Ci\text{^}2\right) \tag{1}$$

的高斯函数对数据点进行函数逼近的拟合方法，跟多项式拟合类比起来，多项式拟合使用的是幂函数，高斯拟合使用的是高斯函数系。上式(1)中：Ai 为归一化系数，Bi 为函数最大值位置，Ci 为函数的幅宽度。

2.1.2. 高斯拟合结果

对长春市 2022 年 3 月 26 日之前未发放蔬菜包时的感染人数使用 Matlab 拟合工具箱进行高斯拟合，拟合后的函数为：

$$f(x) = a_1 * \mathrm{e}^{-\left(\frac{x-b_1}{c_1}\right)^2} + a_2 * \mathrm{e}^{-\left(\frac{x-b_2}{c_2}\right)^2} + a_3 * \mathrm{e}^{-\left(\frac{x-b_3}{c_3}\right)^2} + a_4 * \mathrm{e}^{-\left(\frac{x-b_4}{c_4}\right)^2} \tag{2}$$

其中函数的各个系数如下所示：

$$a_1 = 40200,\ b_1 = 22.18,\ c_1 = 5.092$$
$$a_2 = 1353,\ b_2 = 19.24,\ c_2 = 1.409$$
$$a_3 = -39230,\ b_3 = 22.02,\ c_3 = 4.835$$
$$a_4 = 892,\ b_4 = 9.182,\ c_4 = 0.8733$$

利用拟合出的函数进行反预测，画出拟合后的函数图像，并画出长春市疫情开始时实际感染人数图像，用于二者对比，如下图 1 所示：

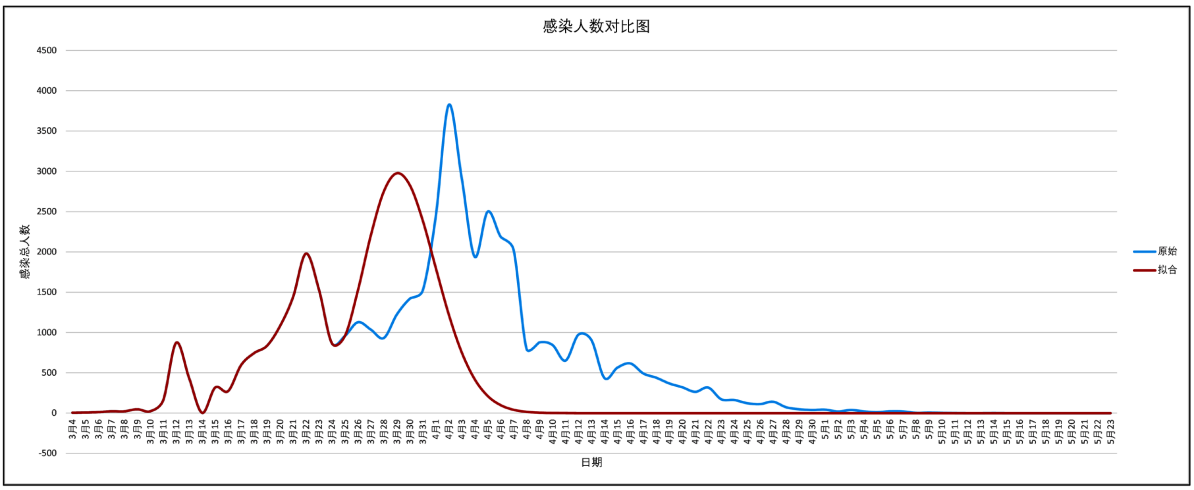

Figure 1. Comparison of the total number of infections
图 1. 感染总人数对比图

红色线为通过高斯插值拟合出的函数曲线，蓝色线为题目所给的实际数据绘制的曲线。由对比图可看出，拟合函数的预测数据的走势与原数据的走势相同，拟合结果具有较高的正确性。由于发放蔬菜包导致人员之间的交流，发生了交叉感染，疫情持续时间增长，感染人员数量的清零时间明显延后，感染人员的最高数值从 2976 增加至 3823，感染人数到达峰值的时间延后。发放蔬菜包时的人员流动导致疫情周期变长，对疫情有着一定的不良影响。

2.2. 算法概述

2.2.1. K-Means 算法概述

K-means 算法是硬聚类算法，是典型的基于原型的目标函数聚类方法的代表，它是数据点到原型的某种距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。K-means 算法以欧式距离作为相似度测度，它是求对应某一初始聚类中心向量 V 最优分类，使得评价指标 J 最小。算法采用误差平方和准则函数作为聚类准则函数。K-means 核心思想为：由用户指定 K 个初始质心(initial centroids)，作为聚类的类别(cluster)，重复迭代直至算法收敛。即以空间中 K 个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果[1]。由于整个长春市划分很多区，所有区的分布图如图 2 所示：

Figure 2. Distribution of districts in Changchun
图 2. 长春市各个区分布图

2.2.2. MST 算法概述

多生成树(MST)是把 IEEE802.1w 的快速生成树(RST)算法扩展而得到的。采用多生成树，能够通过干道(trunks)建立多个生成树，关联 VLANs 到相关的生成树进程，每个生成树进程具备单独于其他进程的拓扑结构；MST 提供了多个数据转发路径和负载均衡，提高了网络容错能力，因为一个进程的故障不会影响其他进程。

MST 性质：假设 $N=(V,\{E\})$ 是一个连通网，U 是顶点集 V 的一个非空子集。若 (u,v) 是一条具有最小权值(代价)的边，其中 $u\in U$，$v\in V-U$，则必存在一颗包含边 (u,v) 的最小生成树。

本题以距离最小为目标建立模型 $D=\sum_{i,j=1}^{n}d_{ij}$，其中 d_{ij} 为 i 投放点到 j 投放点的距离。Prim 算法过程为：假设 $N=(V,\{E\})$ 是连通图，TE 是 N 上最小生成树中边的集合。算法从 $U=\{u_0\}(u_0\in V)$，$TE=\{\}$ 开始，重复执行下述操作：在所有 $u\in U$，$v\in V-U$ 的边 $(u,v)\in E$ 中找一条代价最小的边 (u_0,v_0) 并入集合 TE，同时 v_0 并入 U，直至 $U=V$ 为止。此时 TE 中必有 n-1 条边，则 $T=(V,\{TE\})$ 为 N 的最小生成树。

2.2.3. K-Means 聚类结果

这里只以朝阳区为例，主要使用 K-means 算法对朝阳区的小区进行分簇，假设分为 4 簇[2]，随机在样本集 D_i 中随机选取四个样本，求出最优的均值向量(质心)。以朝阳区为例，首先需要画出朝阳区的小区散点图(其余小区的散点图以及分类图随附件发送)，使用贪心策略寻找最优质心，算法描述如下[3]：

输入：朝阳区小区样本集 $D_i=\{X_i,Y_i\}$；
聚类簇数 $k=4$
过程：
1：从 D 中随机选择 4 个样本作为初始均值向量 μ_i $(i=1, 2, 3, 4)$
2：令 $C_i=[\,](1\le i\le 4)$
3：遍历一遍朝阳区小区的坐标地理位置，画出散点图
4：for $j=1:n$ %% (n 为朝阳区小区数量)
计算朝阳区小区样本 x_j 与各均值向量 μ_i 的距离 $d_{ji}=\left\|x_j-\mu_i\right\|_2$
根据距离最近的均值向量确定 x_j 的簇标记：$\lambda_j=\arg\min_{i\in\{1,2,3,4\}} d_{ji}$
将样本 x_j 划入相应的簇：$C_{\lambda j}=C_{\lambda j}\cup\{x_j\}$
end
for $i=1:4$
计算新均值向量：$\mu_i'=\frac{1}{C_i}\sum_{x\in C_i}x$
if $\mu_i'\neq\mu_i$
将当前均值向量 μ_i 更新为 μ_i'
else
保持当前均值向量不变
end

根据上述算法求解出朝阳区最优的质心坐标，这四个质心坐标将朝阳区的所有小区分成四个簇，每个簇的质心与这个簇内的所有小区坐标距离最小，可以当作物资发放点。如下图 3 所示：

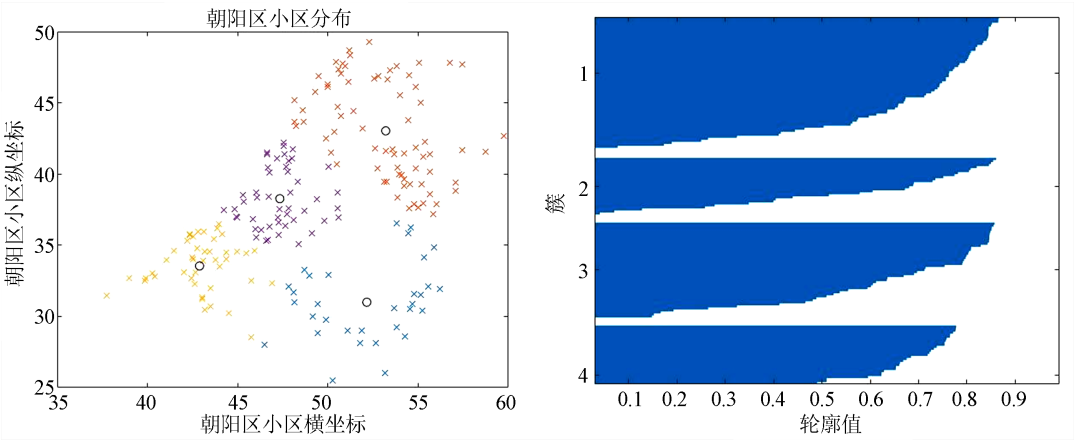

Figure 3. Results of 4 clusters and contour values of clustering effect in Chaoyang District
图 3. 朝阳区小区分 4 簇结果及分簇效果轮廓值

从上图 3 可以看出，将朝阳区分成四簇，从上图 3 的轮廓值图可以看出分簇的效果较好。当分为 6 簇，8 簇时，质心坐标及其轮廓值如下图 4，图 5 所示：

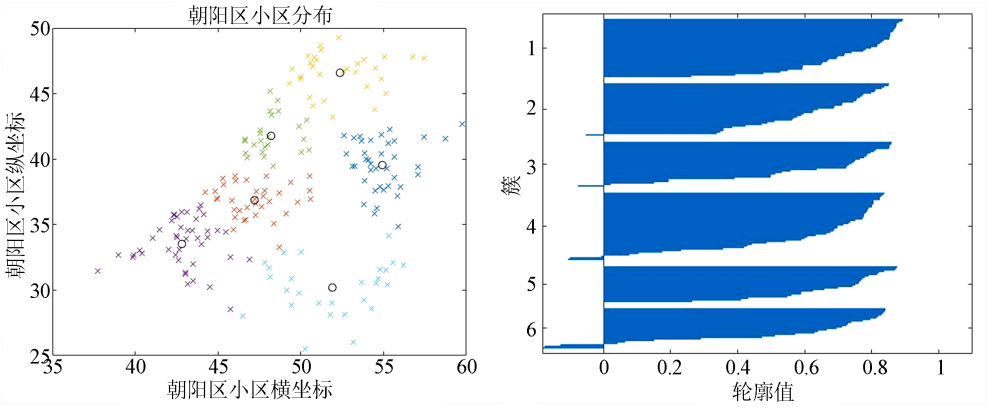

Figure 4. 6-cluster results and contour values of clustering effect in Chaoyang District
图 4. 朝阳区小区分 6 簇结果及分簇效果轮廓值

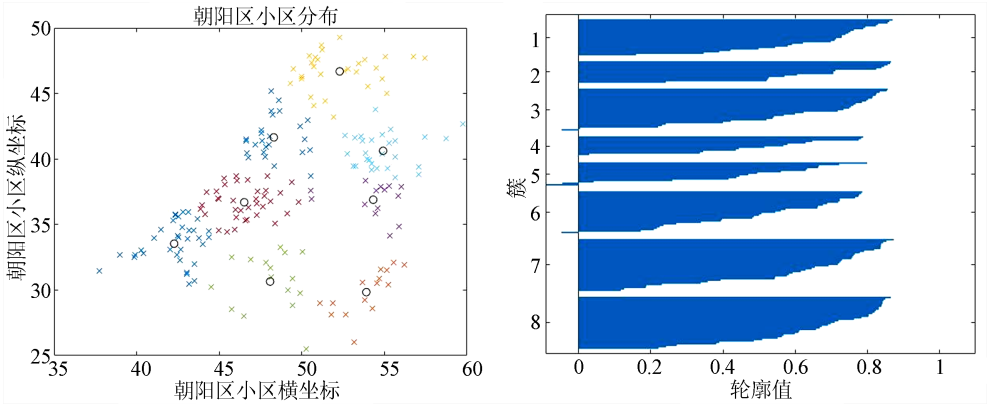

Figure 5. 8-cluster results and contour values of clustering effect in Chaoyang District
图 5. 朝阳区小区分 8 簇结果及分簇效果轮廓值

根据上图 4，图 5 可以看出，当分为 4 簇和 6 簇时，轮廓值较高，最大值都在 0.85 以上，表示分簇的效果比较好，当分为 8 簇，轮廓值在 0.85 以下，分簇效果不好[4]，所以本文各区分簇都以 4 簇为准。当分为 4 簇时，每个质心管理的小区数目太多，可以将分好的 4 簇基础上在利用 K-means 算法进行聚类，k 值选择为 6，结果如下图 6 所示：

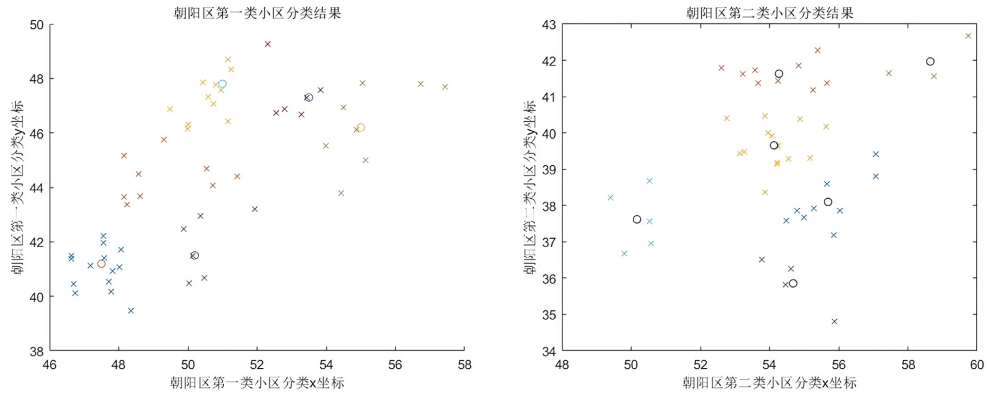

Figure 6. Re-clustering and center of mass of the first and second type of cells in Chaoyang District
图 6. 朝阳区第一类，第二类小区再次分簇及质心示意图

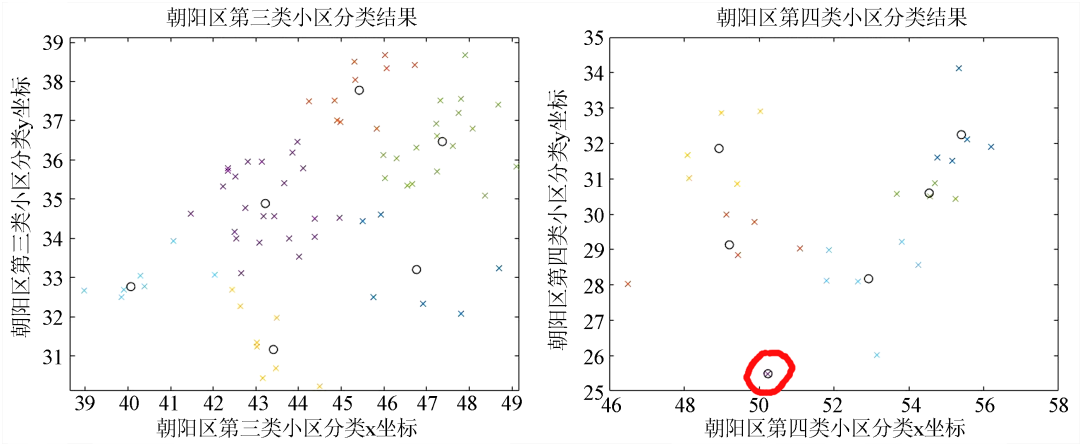

Figure 7. Re-clustering and center of mass of type III and type IV cells in Chaoyang District
图 7. 朝阳区第三类，第四类小区再次分簇及质心示意图

当在朝阳区第一次分簇的基础上再次分簇时，在图 7 中出现了单个小区为一簇的情况，如标记所示，这是 K-means 算法本身的局限性导致的，解决这种问题的方法可以改变 K 值来增加或者减少分簇情况或者多次运行程序来解决[5]。利用 K-means 算法对长春市朝阳区所有小区进行分簇后，得到簇心坐标，可作为合理的蔬菜投放点，簇心坐标和其管理的小区坐标如下表 2 所示：

Table 2. Coordinates of vegetable drop-off locations in Chaoyang District
表 2. 朝阳区蔬菜投放位置坐标

投放点坐标	管辖人数	投放点坐标	管辖人数
(4718.4525, 41.006)	18,356	(54.2741, 41.6301)	25,836
(50.5967, 47.3075)	32,023	(54.1307, 39.6573)	46,080
(55.2697, 46.3343)	26,083	(54.6827, 35.8547)	7879
(53.0388, 47.3953)	21,493	(58.6487, 41.9648)	9189
(49.3073, 44.3646)	32,829	(50.1677, 37.6191)	15,843
(50.4763, 41.8826)	16,100	(46.7658, 33.2057)	18,156
(55.6917, 38.1002)	23,705	(45.4233, 37.7778)	22,856
(43.4108, 31.167)	21,144	(43.2213, 34.8901)	41,742
(55.405, 32.2469)	4228	(48.9307, 31.8573)	50,394
(52.9257, 28.1732)	14,411	(54.5448, 30.595)	11,945
(47.3698, 36.4691)	43,409	(40.0692, 32.7699)	29,655

利用 K-means 算法求解出最佳的物资投放点位置，结合长春市朝阳区各个小区坐标和长春市道路坐标发现，使用 K-means 算法得出的质心坐标并不存在与小区坐标冲突，并且大部分都靠近道路附近，那些不靠近道路坐标的质心点可以在附近择优选择最近的道路作为物资投放点。

根据上述得出的质心坐标，结合 MST 算法，以各个投放点之间的距离为目标，建立投放点之间最小距离为数学模型 $D=\sum_{i,j=1}^{n} d_{ij}$，(其中 d_{ij} 为 i 投放点到 j 投放点的距离)求解出的最佳运输路线结果如图 8 所示：

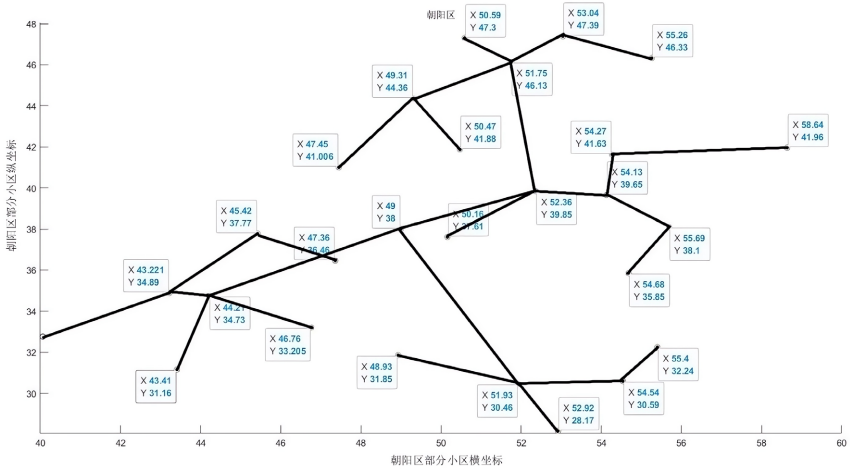

Figure 8. Map of optimal material delivery in Chaoyang District
图 8. 朝阳区最佳物资输送图

为了验证得到的模型的普适性，以长春市其他城区小区坐标为例，下面以长春市宽城区为例，利用上述的模型得出物资投放点坐标，如表 3 所示，并利用 MST 算法得出物资运输路线图，如图 9 所示。

Table 3. Coordinates of vegetable placement locations in Kuan Cheng District
表 3. 宽城区蔬菜投放位置坐标

投放点坐标	管辖人数	投放点坐标	管辖人数
(59.8304, 56.2938)	10,721	(62.6478, 56.6715)	9604
(62.8783, 50.3242)	25,242	(51.0267, 63.6614)	51,734
(59.2782, 51.9657)	17,044	(49.6743, 67.4576)	14,514
(53.8503, 68.7547)	27,832	(46.6612, 62.0177)	17,696
(56.4196, 59.775)	48,204	(61.7611, 64.2048)	16,593
(62.618, 59.5869)	10,440	(58.5887, 65.7671)	41,869
(54.1372, 55.6812)	11,285	(55.3441, 52.2336)	9864
(52.8956, 50.877)	6550	(57.2865, 49.7002)	13,099

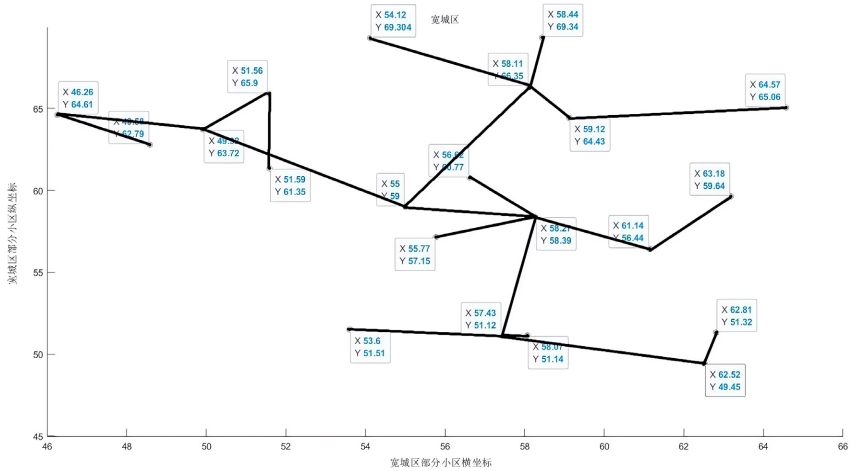

Figure 9. Map of the best material transport in Kuancheng District
图 9. 宽城区最佳物资输送图

从上述求出的质心坐标并结合宽城区的小区坐标以及长春市交通道路坐标，各个质心坐标与宽城区的小区坐标并不存在冲突，通过 K-means 得出的质心坐标和 MST 相结合，画出物资配送路线图，可作为以后发生困难时的运送物资参考路线。

3. 总结

本文以长春市朝阳区为例，鉴于疫情期间的蔬菜物资发放点不合理，发放方式不当容易造成二次传播的问题，首先采用高斯拟合的方法，对假设如果不统一发放蔬菜包的问题进行评价，通过最后拟合出的函数可以看出，疫情期间蔬菜物资发放点不合理，发放方式不当的问题确实加速了病毒的传播。然后以朝阳区为例，对物资投放点不合理的问题，采用 K-means 算法对朝阳区所有小区进行分簇，得到的簇心坐标可以作为理论上的生活物资投放点，并结合 MST 算法，得出物资运输路线图。为了得出模型的普适性，本来以长春市宽城区为例，通过上述模型，同样得出了较为合理的物资投放点坐标以及物资运输路线，为以后发生类似情况时，提供了参考。

参考文献

[1] 徐昊源, 缪鸿志. 基于 K-means 聚类的生鲜自提柜选址及配送方案优化[J]. 物流技术, 2022, 41(11): 50-54.

[2] 黄雨珊, 李钢, 金安楠, 于悦. 社区化新零售末端物流网络的对接与优化——以深圳市盒马鲜生与菜鸟驿站为例[J]. 地理研究, 2021, 40(9): 2542-2557.

[3] 倪卫红, 陈太. 基于聚类-重心法的应急物流配送中心选址[J]. 南京工业大学学报(自然科学版), 2021, 43(2): 255-263.

[4] 梁玥, 陈思, 汤银英. FA-kmeans 算法下面向城乡物流网络优化的网点选址研究[J]. 综合运输, 2021, 43(5): 115-122.

[5] 韩萌. 生鲜农产品城市物流配送中心选址研究[D]: [硕士学位论文]. 兰州: 兰州财经大学, 2018.