“拍照赚钱”任务定价

摘要

本文就“拍照赚钱”任务定价问题，采用聚类分析、回归分析等方法，对目前定价的规律进行了分析，进而建立了任务定价优化模型和任务发布规划模型，给出了新的定价和发布方案。

针对问题一，首先结合 K 均值聚类法和任务地理位置特点，将附件 1 中的任务点按距离分类。从实际的地理位置来看，16 个类的中心都处于较为繁华或者交通便利的城市中心位置，而任务点到相应中心的距离和定价有明显的线性相关性。然后用任务点到相应中心的距离对平均价格进行回归，找出了当前定价方案总的定价规律。最后，对当前任务完成情况进行了统计分析，总结出完成情况不佳的主要是因为定价没有利用会员信息和体现出不同类之间的差别。

针对问题二，首先在各个类的内部，用任务点到相应中心的距离和会员信息对价格进行了回归，给出了各类的初步定价。然后用完成度函数对任务完成情况进行量化处理，以平均完成度最大为目标建立了任务定价优化模型一，求解模型。最后，将新的定价方案和原来的方案进行了比较，结果显示新的方案的完成度有显著的提高，并且各类的平均定价也更加合理。

针对问题三， 首先依据距离远近给出了任务打包方案，在模型一的基础上建立了打包任务点定价模型二，并给出了打包后的定价方案。然后，以任务成本最小为目标，建立了任务发布规划模型三，并用 lingo 对模型求解。然后根据模型三的解，结合“会员竞争力函数”，给出了任务发布的具体方案。最后，将原方案、模型一定价方案和模型三给出的定价方案三者进行了比较，结果表明模型三给出的方案的成本更低且完成度最高。

针对问题四，首先重新对附件三中的任务位置进行分析，结合问题一中的分类结果，可以将附件三中的任务分为三类。然后利用模型二和模型三对新的任务进行了打包、定价和发布。

最后，对模型的优缺点进行了评价，并对模型的应用进行了推广，具有一定的现实意义

关键词：0-1 规划 聚类分析 动态加权函数 完成度函数 会员竞争力函数

1. 问题重述

随着时代的发展，大数据时代已经到来，移动互联网早已是社会中常见的一种网络模式。在大势所趋的移动互联网时代，产生了一种自助式服务模式——“拍照赚钱”。所谓“拍照赚钱”，就是用户去下载 APP，注册成为 APP 的会员，然后从 APP 上领取需要拍照的任务，然后赚取 APP 对任务所标定的酬金。

这种新兴的自助式劳务众包平台，不仅为企业提供各种商业检查和信息搜集，而且也为有相应工作需求的人提供一个比较方便的平台。这种新兴的市场调查方式与传统的市场调查方式相比，一方面可以大大节省调查成本，另一方面也有效地保证了数据调查的真实性，缩短了调查的周期。

所以，该 APP 是这个自助式劳务众包平台运行的核心，而对于这个 APP，研究其定价又成为一个核心任务。

题目要求根据附件中的数据，对不同任务点的不同定价进行分析和重新制定新的定价方案，主要需要解决以下问题：

1.研究附件一中项目的任务定价规律，分析任务未完成的原因。

2.为附件一中的项目设计新的任务定价方案，并和原方案进行比较。

3.实际情况下，多个任务可能因为位置比较集中，导致用户会争相选择，一种考虑是将这些任务联合在一起打包发布。在这种考虑下，如何修改前面的定价模型，对最终的任务完成情况又有什么影响？

4.对附件三中的新项目给出你的任务定价方案，并评价该方案的实施效果。

2. 问题的分析

“拍照赚钱”这种在移动互联网下的自助式服务模式，带来了不少好处和方便。同时也衍生出一系列的问题，对于这些问题，分析如下。

2.1 问题一的分析

对于问题一，需要解决的是根据附录一中任务点的信息数据找出其标价规律，并找出影响任务完成度的原因。首先将所给数据的经纬度做一定的坐标变换，画出全局任务点分布图。我们首先想到的是按照距离做 K 均值聚类分析，将 835 个任务点进行分类，然后分总体标价规律与各类的标价规律两种分析思路，分别进行线性回归的方法找出任务点标价规律。

2.2 问题二的分析

对于问题二，我们需要重新制定一个定价方案，并与原方案进行比较。在这里我们给出了一种新的定义——完成度，顾名思义它表示完成的任务数在总任务点中所占的比率。因此，为了找到一个更优的定价方案，我们需要建立一个优化模型，使总的完成度最大化。当然，与原来的定价方案相比，其完成情况肯定得到了明显的改善。

2.3 问题三的分析

对于问题三，我们需要进行三个步骤：（1）对任务点进行打包处理；（2）对打包后的任务包重新定价；（3）把所有任务（包括任务包）根据会员的竞争力顺序分发出去。首先打包处理，能够进行打包的任务点之间肯定不会距离很远，因此我们需要分别计算出各区域内任务点间的距离矩阵，设置一个合适的阈值，对两任务点距离小于该阈值的两任务进行合并成一个任务包，依此类推下去，将所有任务打包完成；其次重新定价，当然，与其它任务的距离均未在阈值之内的任务点，成为单个任务点包，其定价依然等于问题二中的价格，剩下的任务包需要重新定价；最后，把任务分发下去，作为任务定价方，在发布的任务能够被会员完成的前提下，当然希望成本越低越好，但也并不能将任务过于集中发布给一些会员，因为他们都是有任务量限额的，并且每个任务有发布给哪个会员的选择，在这几方面综合考虑下，我们可以考虑以成本最小为目标函数，各会员的任务量限额为约束条件，建立一个 0-1 线性规划模型。

2.4 问题四的分析

对于问题四，附录三给出了一系列新的任务点位置信息，根据问题一的分类结果，我们可以把他们分进对应所属的类中，然后就可以根据问题三所给出的打包定价发布的方法对这些任务进行处理，最后还可以应用任务的完成度函数对处理结果进行评价。

3. 假设与符号说明

3.1 假设

（1）不考虑 GPS 与地表坐标转换之间的误差；
（2）假设任务点所在区域内，会员去完成任务时，对距离的定义只考虑到地理距

离，不考虑其他因素（如交通、天气情况）的影响；

（3）假设所有任务完成的难度相同，所花精力值相同。即不存在任务本身性质的影响；

（4）假设会员不会共享任务信息。

3.2 符号说明

（1）x_{ij}：表示第 i 类的第 j 个任务点到相应聚类中心的距离

（2）p_{ij}：表示第 i 个子区域的第 j 个任务点的标价

（3）$I_{ij}=\begin{cases}0\\1\end{cases}$，当$I_{ij}=1$时，表示第 i 个区域的第 j 个任务被完成；当$I_{ij}=0$时，表示第 i 个区域的第 j 个任务未被完成

（4）d_{ij}:第 i 个区域的第 j 个任务点到聚类中心的距离

（5）$\overline{p_i}$：表示区域平均价格，其计算公式为：

$$\overline{p_i}=\frac{1}{m_i}\sum_{j=1}^{m_i}p_{ij}\text{，}\ \mathrm{i}=1,2,\cdots,16$$

（6）d_i：表示第 i 类聚类的半径

（7）k_i：表示第 i 类的系数（比例）

（8）p'_{ij}：表示第 i 类第 j 个任务的拟合值

（9）d_{ijk}：表示第 i 类中的第 k、j 两个任务点的距离

（10）a:表示第 k 个会员的任务量的限制

（11）r_k：表示第 k 个会员的信誉值

（12）y_{jk}：表示第 k 个会员对第 j 个任务点的竞争力

（13）d_{jk}：表示第 k 个会员到达第 j 个任务点的距离

（14）k_{ij}:表示在第 i 个区域内以第 j 个任务点为圆心到第 j 个点的距离小于k_0的会员的个数（k_0的设定未定，初值为相应的聚类区域的半径）

（15）ρ_k：附件 2 中第 k 个会员的最大限度

4. 模型的准备

4.1 经纬度与平面坐标系的转换

问题一要求我们研究附件一项目中的任务定价规律，如果我们想要对题目给出数据做进一步的分析和处理，那么我们首先要做的就是把题目所给的项目的经纬度转化为我们熟知的平面二维坐标。其转化方法为：利用经纬度与平面二维坐标的转换公式，将题目所给的经纬度转化为平面二维坐标。其具体纬度计算公式为[1]

$$\mathrm{S}=\frac{\cos \mathrm{B}\times(\mathrm{R}-\mathrm{B}\times(\mathrm{R}-\mathrm{r})/90)\times 2\pi}{360} \qquad (1)$$

经度计算公式为：

$$\mathrm{S}=\frac{[\cos B_1\times(R-B_1\times(R-r)/90)+\cos B_2\times(R-B_2\times(R-r)/90)]\times\pi}{360} \qquad (2)$$

其中，B为计算点所在的纬度（取到度），R为迟到半径，r为极轴半径，取$\mathrm{R}=6738137\mathrm{m}$，$\mathrm{r}=6356752\mathrm{m}$，$B_1$和$B_2$为两点的纬度（取到度），$(R-r)/90=(6378137-6356752)/90=237.6111$。

4.2 K 均值聚类分析[2]

聚类分析就是根据“物以类聚”的道理，对样本或指标进行分类的一种多元统计分析方法。讨论的对象是许多样本，合理地按各个样本的特性进行分类，而且没有任何模式可以参考。聚类分析首先根据样本的多个观测指标，具体找出可以度量样品或者指标间相似程度的一些统计量，然后利用这些统计量把样品或指标归类。聚类分析的目的就是把分类对象按照一定的规则划分成若干类，对类的数目和结构没有必要做任何假设。其中，K 均值聚类法是动态聚类法中的一个典型代表，它是由麦奎因提出并命名的一种算法。其基本步骤为：

（1）随机选择 K 个样品，或将所有的样本分成 k 个初始类，然后把这 k 个类的均值当作初始凝聚点。

（2）把除了凝聚点之外的其他样品逐个归类，将所有样品归入凝聚点离它最近的那一个类别（一般采用欧式距离），那么该类的凝聚点就更新为这一类目前的均值， 直到所有的样品都已经归类。

（3）重复步骤（2），直到所有样品不能再分配为止。

4.3 动态加权函数

借鉴S型分布函数，给出一个完成度函数，其具体定义如下[3]：如果某项指标$\mathrm{x_j}$对于综合评价结果影响大约是随着类别$p_k(k=1,2,\cdots,K)$的增加而增加的过程，呈一条S型曲线，那么此时对指标$\mathrm{x_j}$的动态加权函数可以设定为S型分布函数。即

$$w_j(x)=\begin{cases} 2\left(\frac{x-a_1^{(j)}}{b_K^{(j)}-a_1^{(j)}}\right)^2,当a_1^{(j)}\le x\le c时, \\ 1-2\left(\frac{x-a_1^{(j)}}{b_K^{(j)}-a_1^{(j)}}\right)^2,当c<x\le b_K^{(j)}时, \end{cases}$$

其中，参数$\mathrm{c}=\frac{1}{2}\left(a_1^{(j)}+b_K^{(j)}\right)$，且$w_j(c)=0.5(1\le j\le m)$。

4.4 会员竞争力函数

实际情况下，有很多个任务可能会因为位置比较集中，导致会员很可能会争相竞争选择，所以我们定义一个会员竞争力函数。其具体定义如下：

对第 j 项任务而言，会员竞争力为：

$$y_{jk}=\frac{r_k}{d_{jk}}$$

此函数就称为会员竞争力函数。

5. 模型的建立与求解

5.1 当前定价方案规律分析

5.1.1 经纬度与平面坐标之间的转换

把题目给定的附件 1 中的经纬度分别带入 4.1 节中的公式（1）和公式（2），则可以计算出相应的平面坐标，转换后附件 1 中的数据分布如下图 1 和图 2，其中图 1 为整体图像，图 2 为局部放大图形。

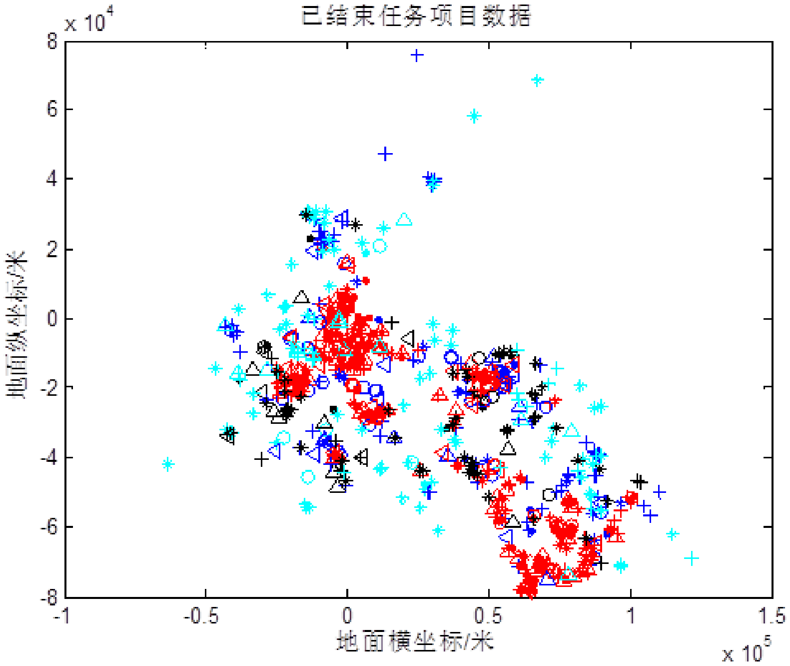

图 1 经纬度转换为平面坐标（总体）

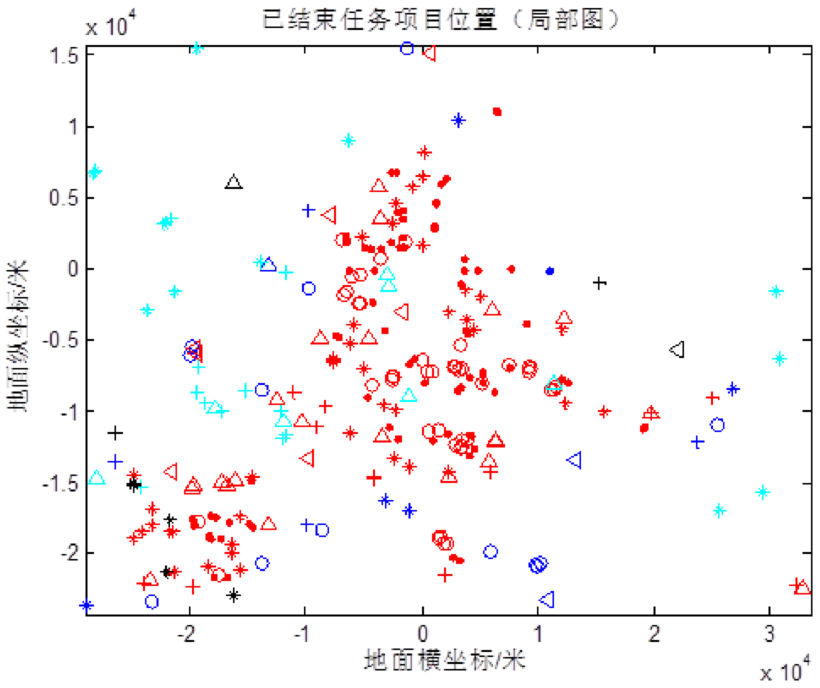

图 2 经纬度转换为平面坐标（局部）

图 1 和图 2 中，红色标记的点代表标价最低，红色的点中，按照圆圈、实心点、星号、三角、加号、左向三角的顺，价格依次上升。由图 1 和图 2 可以知道，任务点有一定的集群现象，而且，我们可以从图中很明显看出当前的定价方案和任务点到相应中心的距离有明显的相关性。总体上来说，定价低的任务点有集中的趋势，因此可以考虑按距离对任务点聚类。

5.1.2 任务点聚类

我们通过将数据可视化后，可以明显的观察到数据点存在集群现象，通过聚类分析后，我们可以得到任务标价与任务到聚类中心的距离的相关关系。与此同时，我们先判断所有聚类分析的聚类中心是否准确，其根据是通过任务标价与任务到聚类中心的距离的相关关系的大小。我们首先将小价格的位置进行聚类，提取出16个聚类中心，然后对16个聚类中心生成的数据点到聚类中心的距离，将这些点与任务标价进行相关性分析，得到皮尔逊系数，判断数据点到聚类中心距离与任务标价之间的关系。结果如下表1所示。

表1 聚类中心点坐标

Obs	横坐标	纵坐标
1	65125.411	-76152.88626
2	83915.3473	-69908.20295
3	-19263.67515	-17761.07632
4	73888.104	-23598.37145
5	100472.567	-51675.49819
6	6218.027237	-28701.43544
7	-8369.966866	21313.5799
8	-1424.956676	1884.486627
9	38662.30315	-28473.09132
10	420.1983189	-7191.207417
11	51615.66667	-42340.36069
12	11066.77994	-8428.49249
13	-4178.947956	-39599.0757
14	53870.34262	-58615.92726
15	48182.38193	-16135.92294
16	74545.37668	-59938.65005

由表1可以知道，我们将16个类的聚类中心由经纬度转化为平面坐标后的二位坐标系的位置。

这16个坐标点，有一定的实际意义，都是繁华的商业中心或者交通便利的市中心。比如第4个点为广东省东莞市，是处于市中心。

根据我们的分析和研究，我们发现，当该任务点偏离聚类中心的程度越大，则该任务点的定价也就越高，即我们有理由认为该方案呈现出按照距离的远近来定价的趋势，具体情况如图3所示。图4为局部聚类点的结果显示图。

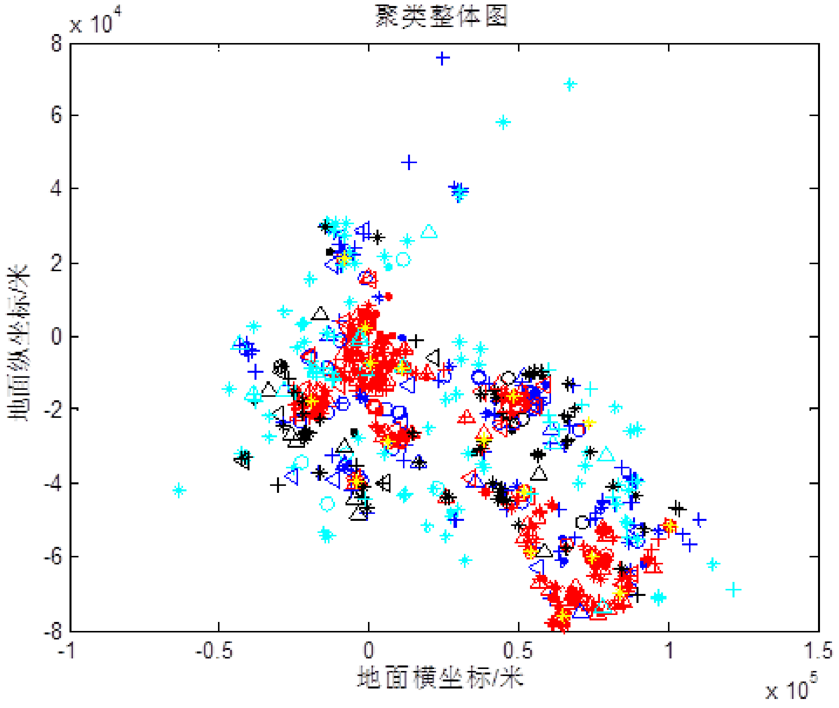

图 3 整体聚类图

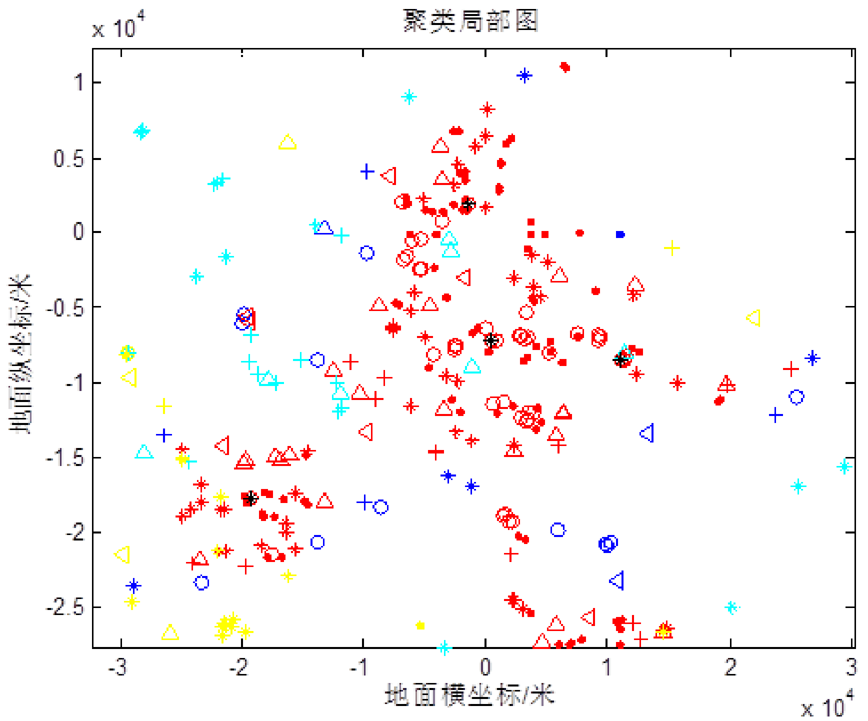

图 4 局部聚类图

而从实际位置来看，这些聚类中心大致都是在繁华地带或者交通便利之处，而且我们还可以从图中知道，任务标价与数据点到中心距离存在一定的相关关系，通过图像可知：数据点到聚类中心距离越远时，任务标价越高。

5.1.3 总结定价规律

（1）总体规律

通过图 3 和图 4 可以知道，任务点的分布大概都是在市中心等繁华地带或者交通便利的城市，而且呈现出集群分布的现象。由此可知，任务点到聚类中心的距离与该任务点的定价有一定的线性关系，即呈现出按距离的远近来定价的趋势。从图 5 回归的结果来看，任务平均价格相对于任务点到类中心的距离有显著的线性关系。回归方程为：

$$y = 0.7466x + 65.751,$$

$R^2 = 0.96$, F值为 6.47，对应的 P 值为 0.0001<0.01，因此回归具有显著意义。

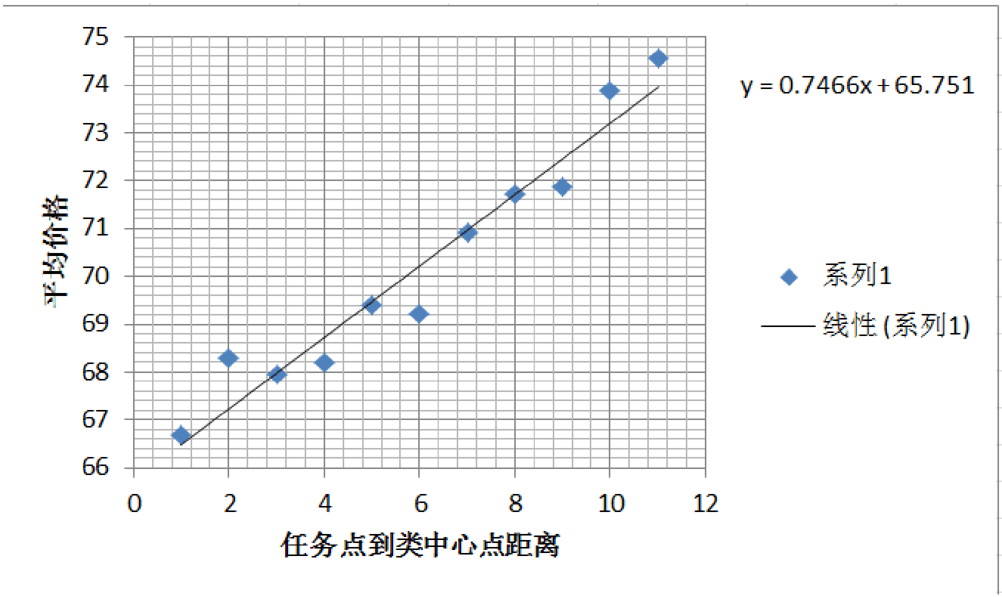

图 5 平均价格与任务点到类中心点距离

（2）各类的（局部）规律

在类的内部，可以发现相似的线性规律，不同类的线性回归系数有所不同而已，因此不再赘述。

5.1.4 未完成原因

由上述我们对已有定价规律方案的分析，通过对上述数据的处理以及对原有的定价方案的特点和规律的总结，我们可以发现任务未完成可能有以下几个原因：

1. 原定价方案是纯粹按照距离定价，不同的任务点之间的定价差别太小，而且没

有考虑任务的密集程度的影响等；

2. 原定价方案没有考虑会员的信息，没有将任务和会员进行合理的分配；

3. 原定价方案没有对任务进行整合处理，比如将比较临近的几个任务打包处理等。

5.2 任务定价优化模型

5.2.1 定价影响因素分析

对一个商品定价，会受到很多因素的影响，比如在本题中，题目要求我们对附件一中的项目设计一套新的定价方案，那么首先我们需要考虑的就是会对定价产生的影响的因素有哪些。

1. 距离对定价有直接的影响。此处的距离有任务点到聚类中心的距离、会员到聚类中心的距离以及会员到任务点的距离；

2. 会员信息也会对定价产生一定程度的影响，比如以任务点为中心、任务点到聚类中心的最大距离为半径围成的圆的范围内的会员个数，会员自身的信誉值等因素都有可能会对我们的定价产生一定的影响。因此，我们接下来需要做的就是找出这些会影响定价的因素与价格之间到底是一个什么样的关系。

5.2.2 回归初步定价

（1）多元线性回归模型。我们给出多元线性回归模型的一般形式，其一般形式如下[4]：

设随机变量y与一般变量 x_1，x_2，⋯，x_p的线性回归模型为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{1}$$

(1)式中，β_0，β_1，β_2，…，β_p是$p+1$个未知参数，β_0称为回归常数，β_1，β_2，…，β_p，称为回归系数。y称为被解释变量（因变量），x_1，x_2，…，x_p是p个可以精确测量病控制的一般变量，称为解释变量（自变量）。$p = 1$时，(1)式即为一元线性回归模型式；$p \geq 2$时我们就称(1)式为多元线性回归模型。ε是随机误差，与一元线性回归一样，对随机误差项，我们常常假定

$$\begin{cases} E(\varepsilon) = 0 \\ Var(\varepsilon) = \sigma^2 \end{cases} \tag{2}$$

称

$$E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p \tag{3}$$

为理论回归方程。

对一个实际问题，如果我们获得n组观测数据

$(x_{i1}, x_{i2}, \ldots, x_{ip}; y_i)(i=1,2,\ldots,n)$，则(1)式线性回归模型可表示为：

$$\begin{cases} y_1=\beta_0+\beta_1 x_{11}+\beta_2 x_{12}+\cdots+\beta_p x_{1p}+\varepsilon_1 \\ y_2=\beta_0+\beta_1 x_{21}+\beta_2 x_{22}+\cdots+\beta_p x_{2p}+\varepsilon_2 \\ \ldots \\ y_n=\beta_0+\beta_1 x_{n1}+\beta_2 x_{n2}+\cdots+\beta_p x_{np}+\varepsilon_n \end{cases} \tag{4}$$

写成矩阵形式为：

$$y=X\beta+\varepsilon \tag{5}$$

(5)式中

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} \qquad X=\begin{bmatrix} 1\ x_{11}\ x_{12} \ldots x_{1p} \\ 1\ x_{11}\ x_{12} \ldots x_{1p} \\ \cdots \\ 1\ x_{n1}\ x_{n2} \ldots x_{np} \end{bmatrix}$$

$$\beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{bmatrix} \qquad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix} \tag{6}$$

X 是一个$n\times(p+1)$阶矩阵，称为回归设计矩阵或者资料矩阵。在实验设计中，X 的元素是预先设定并可以控制的，人的主观因素可作用其中，因而称 X 为设计矩阵。

（2）回归初步定价。通过 5.2.1 对定价的影响因素的分析，我们把价格（P）设为因变量p_{ij}，把会员与聚类中心的距离设为自变量x_1，把以任务点为中心、任务点到聚类中心的最大距离为半径围成的圆的范围内的会员个数设为自变量x_2，做多元线性回归，然后根据这个多元线性回归的方程，对各类的任务定价做初步的估计。对于某个类里面的任务点，我们在做多元线性回归之前，应该先对p_{ij}和x_1以及x_2的相关性做进一步的检验，这一个步骤由 SPSS 实现，其相关性检验结果如下表 2、表 3：

表2 相关性

		任务标价pij	distance'
任务标价pij	Pearson 相关性	1	.228**
	显著性（双侧）		.000
	N	835	828
distance'	Pearson 相关性	.228**	1
	显著性（双侧）	.000	
	N	828	828

**. 在 .01 水平（双侧）上显著相关。

表3 相关性

		任务标价pij	n'
任务标价pij	Pearson 相关性	1	-.109**
	显著性（双侧）		.002
	N	835	835
n'	Pearson 相关性	-.109**	1
	显著性（双侧）	.002	
	N	835	835

**. 在 .01 水平（双侧）上显著相关。

由表 2 和表 3 可知，任务标价 y 与会员到聚类中心的距离x_1的 Pearson 相关性为 0.228，说明任务标价 y 与会员到聚类中心的距离x_1存在一定的正相关。y 与x_2的 Pearson 相关性为-0.109，说明任务标价 y 与x_2存在一定的负相关。任务标价与会员到聚类中心的显著性水平为 0.000 小于给定的显著性水平$\alpha = 0.05$，任务标价 y 与x_2的显著性水平为 0.002 小于给定的显著性水平$\alpha = 0.05$，所以其相关系数具有统计学意义，说明二者具有一定的相关性。用多元线性回归的模型对第 i 类任务进行初步定价：

$$\mathrm{p}_{ij} = \beta_1 x_1 + \beta_2 x_2 + \beta_3,$$

其中，我们选取第 3、6 类稍加分析和说明，见附录。对于第 3 类，其 F 值为 2.73，R 方为 0.1177，三个参数β_1、β_2、β_3分别为 2.99377、2.32364、74、94559，则第 3 类的线性回归模型为：

$$\mathrm{p}_{3j} = 2.99377x_1 + 2.32364x_2 + 74、94559,$$

对于第 6 类，其 F 值为 0.76，R 方为 0.0342，三个参数β_1、β_2、β_3分别为 0.21542、0.94122、72.26086，则第 6 类的线性回归模型为：

$$p_{6j}=0.21542x_1+0.94122x_2+72.26086.$$

其余回归模型用类似的方法得到,具体见附录六。

5.2.3 任务定价优化模型一

题目要求我们对附件一中的项目重新定价，在这个要求之下，我们给出一个任务定价优化模型。

1. 定义完成度函数

首先，定义一个完成度函数$w\left(\frac{p'_{ij}}{p_{ij}}\right)$，当任务是完成了的时候，即任务完成情况是“1”时，其具体的计算公式为

$$w\left(\frac{p'_{ij}}{p_{ij}}\right)=\begin{cases}2x^2，当\ 0\le x\le\frac{1}{2}\\1-2(x-1)^2，当\frac{1}{2}<x\le 1\\1+2(x-1)^2，当\ 1<x\le\frac{3}{2}\\2-2(x-2)^2，当\frac{3}{2}<x\le 2\end{cases}\tag{1}$$

当任务完成情况为“0”时，

$$w_1\left(\frac{p'_{ij}}{p_{ij}}\right)=w\left(\frac{p'_{ij}}{p_{ij}}\right)-1\tag{2}$$

2. 任务定价优化模型一

以平均完成度最大为目标，建立任务定价优化模型如下：

$$\max_{k_i}=\sum_{i=1}^{16}\sum_{j=1}^{m_i}k_i w\left(\frac{p'_{ij}}{p_{ij}}\right)$$

$$\text{s.t.}\begin{cases}0\le k_i\le 1\\\sum_{i=1}^{16}k_i=1\\0\le p'_{ij}\le 2p_{ij}\end{cases}\tag{3}$$

3. 任务最终定价

若 c 为总成本，则

$$c=\sum_{i=1}^{16}\sum_{j=1}^{m_i}p_{ij}=\sum_{i=1}^{16}\left(k_i\sum_{j=1}^{m_i}p'_{ij}\right),\tag{4}$$

$$ck_i = \sum_{j=1}^{m_i} p'_{ij} \quad , \tag{5}$$

则最终的定价为：

$$p''_{ij} = ck_i \frac{p'_{ij}}{\sum_{j=1}^{m_i} p'_{ij}} \quad . \tag{6}$$

5.2.4 模型求解与结果分析

模型的求解步骤如下：

Step1: **计算个任务点的完成度**。根据上述的公式(1)，计算出每一个任务点的完成度函数值；

Step2: **用贪婪算法对模型求解**。这个是一个多元优化问题，求解比较困难，因此采用贪婪算法对问题进行求解，程序见附录 4。结果见下表

表 4 原方案与新方案的k_i

	原方案	新方案		原方案	新方案
k1	0.118433	0.199282	k9	0.069788	0.001348
k2	0.10138	0.156647	k10	0.045682	0.011272
k3	0.057145	0.028613	k11	0.060204	0.032292
k4	0.023742	0.022475	k12	0.039278	0.073766
k5	0.104976	0.178438	k13	0.093287	0.114468
k6	0.057605	0.050989	k14	0.061634	0.034343
k7	0.081503	0.030626	k15	0.019098	0.006156
k8	0.054286	0.0483	k16	0.01196	0.010985

平均完成度　　原来：62.48，新的 72.80

从表的结果来看，原方案的k_i值分布比较均匀，相互的差异比较小，而新方案的k_i值之间相互差异更大一些。

5.2.5 与原方案的比较

（1）平均完成度

原来方案的平均完成度为 62.48，新方案的平均完成度为 72.80，明显新方案的完成度有显著的提高。

(2) 类平均价格

原方案的k_i值分布比较均匀，相互的差异比较小，而新方案的k_i值之间相互差异更大一些。这说明新的定价方案更能反映出类之间从差别。

通过计算，得到 16 个类原始定价的类平均值和模拟的类平均价格，再利用每一类里面的任务点个数，我们便可以求出总的成本价格。相关的计算结果如表 5 所示：

表 5 类平均价格

任务点个数	原定价均值	新定价均值	打包定价均值
101	67.6634	80.8631613	78．30034
88	66.4773	86.728971	74．0024
44	74.9432	59.94318	97.5008
19	72.1053	85.10526	88.0007
86	70.436	57.43604	60．5496
46	72.2609	62.26087	82.0014
67	70.194	87.19402	94.998
46	68.0978	78.09782	74.4988
61	66.0164	54.01638	89.0004
37	71.2432	59.243245	82.9984
51	68.1176	79.5276	82.9976
32	70.8281	63.828	76.4992
78	69.0128	84.0128	79.9984
53	67.1038	89.1037	76.5014
16	68.875	81.875	91.1102
10	69.0128	79.4	80.128

由表 5 可以知道，原定价的均值普遍低于新定价的均值，而且打包后的均值波动范围更大，说明其覆盖率更广。原来的平均完成度 为：62.48，新的为 72.80，打包之后的任务完成度为 85.7。所以，综上所述，优化之后的定价方案比原给定的定价方案好。

5.3 任务发布规划模型

5.3.1 任务打包（不超过四个点打包）

首先我们考虑打包，只有足够近的任务才能够进行打包，那么我们可以设置一个阈值，使得：相邻两个任务可以进行打包。那么我们就可以把打包好的任务点看作是一个新的任务点。合并后，把离聚类中心更远的点的坐标作为合并后的任务点的坐标（打包后两到四个点可以近似看作是一个点），接下来进行下一步骤的操作。

然后我们考虑打包过程中的循环过程，当两个任务点进行打包之后，其阈值

还能继续进行打包的，那么继续和第三个较近的任务点进行再一次的打包；如果还没有超过打包的阈值，那么可以继续上述步骤直至两个任务点之间的距离超过打包的阈值，那么这样就对原来的任务点进行了一个压缩。与此同时，这种打包的方式也对我们的定价模型造成了一定的影响。打包的任务点的价格比原来的每一个的单价要高，但是比原来的所有在打包内的任务点的平均价格要低。因此，我们需要对打包后的任务设定一个新的定价方案或者要对之前的定价方案做一定的修改和优化。

对此，我们给定一个算法，算法具体步骤如下：

1. 先任意选取两个任务点，如果它们之间的距离少于 k 米，那么将这两个任务点打包；
2. 将这两个任务点打包之后，选取周围较近的第三个点。如果两个任务点的中点，到第三个任务点的距离小于 k 米，那么再进行打包；反之，则不进行打包；
3. 执行第二步，直到不再进行打包为止；
4. 在本文中，取$\mathrm{k}=500$。

注意：每次打包后产生的新的任务点的位置，为原来的两点连线的中点。

5.3.2 打包任务定价模型二

我们参考出租车拼车的规则，设计一个打包的定价模型，从而使给出的新的定价方案更加合理。对于任务打包的定价，我们给出这样的定价方案： 首先，我们考虑一下没有经过打包的点，即未打包的任务点，我们依然按照问题二中的我们给定的定价方案进行定价。接下来，我们考虑需要打包的任务点，在这一个过程中，我们需要设置一个打包的距离阈值，当两个或者多个任务点超过这个距离阈值时，就不能被打包。

1. **打包权重函数。**如果有x项任务要被打包，则打包的定价满足这样的函数关系：

$$\mathrm{f}(x)=e^{\lambda(1-x)},$$

其中 x 为需要被打包的任务数，λ为一个影响参数，而且λ的取值不限，我们在题目中设定的打包权重函数$\mathrm{f}(x)=e^{\lambda(1-x)}$，在本文中取$\lambda=0.15$，即打包函数为

$$\mathrm{f}(x)=e^{0.15(1-x)}.$$

2. **打包任务定价模型二**

最终的定价模型为：

假设各个任务点不被打包时的定价分别为：p_1，p_2，…，新的定价符号为：

p(1)，p(2)，…，则

如果有 1 个任务点未被打包，则$\mathrm{p}(1)$按照问题二中给出的定价模型定价；

如果有 k 个任务点被打包，则$\mathrm{p}(k)$按照$\mathrm{p}(k)=\sum_{k=1}^{4} p_k \times f(k)$的方式定价。

5.3.3 任务发布规划模型的建立

在任务发布环节，我们需要考虑：会员与任务点（包括打包点）的距离、会员的预定任务限额、预定任务开始时间及会员的信誉值。在这个环节，我们必须制定一个发布的模型。

首先，我们介绍一下 0-1 整数规划的模型：如果整数线性规划问题的所有决策变量x_i仅限于取 0 或者 1 这两个数值，则称此类问题为 0-1 线性整数规划，简称为 0-1 规划。变量x_i被称为 0-1 变量，或者是二进制变量，其变量取值的约束可以变为$x_i=0$，或者$x_i=1$, 等价于$x_i \le 1$和$x_i \ge 0$，而且为整数。那么，0-1 规划的一般模型为

$$\max(min)z=\sum_{j=1}^{n} c_j x_j \text{ ,}$$

$$\mathrm{s.t.}\begin{cases}\displaystyle\sum_{j=1}^{n} a_{ij}x_j \le (=,\ge) b_i \quad (i=1,2,\cdots,m),\\ x_j = 0 \text{ 或 } 1 \ \ (j=1,2,\cdots,n).\end{cases}$$

先设 p_{ij}表示第 i 个子区域的第 j 个任务点的标价，然后设：

$$x_{ijk}=\begin{cases}1\text{，将第}i\text{累第}j\text{项任务发布给第}k\text{个会员}\\0\text{，否则不将此项任务发布给这个会员}\end{cases},$$

根据上述假设，我们以成本最低为目标建立规划模型：

$$\min z=\sum_{i=1}^{16}\sum_{j=1}^{m_i} p_{ij}x_{ijk} \text{ ,}$$

$$\mathrm{s.t.}\sum_{i=1}^{16}\sum_{j=1}^{m_i} x_{ijk} \le a_k,$$

其中，a_k表示第 k 个会员的任务量的限制，z 表示成本。

5.3.4 任务打包的发布模型及求解

（1）打包完成度函数

经过我们对上面的完成度函数的计算，我们发现该模型需要改进，即该模型需要有一个权重，此权重的定义式为：（是给出的定义式，不用证明，直接使用）

$$\mathrm{f}(k)=\frac{4}{\pi}\arctan[tan(1)k-tan(1)+1]$$

且$1\leq \mathrm{f}(k)<2$。则，给出一个新的完成度函数

$$\mathrm{w}_2\left(\frac{p'_{ij}}{p_{ij}}\right)=\begin{cases}\mathrm{w}\left(\frac{p'_{ij}}{p_{ij}}\right)f(\mathrm{k})\text{，当完成情况为“1”时}\\ \left[\mathrm{w}\left(\frac{p'_{ij}}{p_{ij}}\right)+1\right]f(k)\text{，当完成情况为“0”时}\end{cases}$$

（2）任务打包的发布模型及求解

根据题目要求，我们先设$x_{ijk}=\begin{cases}1\text{，将第 i 类第 j 项任务发布给第 k 个用户}\\0\text{，将第 i 类第 j 项任务发布给第 k 个用户}\end{cases}$，

建立规划模型：

$$\min\sum_{i=1}^{16}\sum_{j=1}^{671}\sum_{k=1}^{1877}p_{ij}x_{ijk}\ ,$$

$$s.t.\begin{cases}\sum_{i=1}^{16}\sum_{j=1}^{671}\sum_{k=1}^{1877}x_{ijk}=671\\ x_{ijk}=\begin{cases}1\\0\end{cases}\\ \sum_{i=1}^{16}\sum_{j=1}^{671}x_{ijk}\leq\rho_k\end{cases}$$

我们运用 LINGO 中的 0-1 整数规划的思想，将上述模型进行求解。首先我们可以将第一个小类的任务数和会员数提取出来，然后根据整数规划的思想编出程序。具体的程序结果如下图 6：

```
Global optimal solution found.
Objective value:                              5070.000
Objective bound:                              5070.000
Infeasibilities:                              0.000000
Extended solver steps:                               0
Total solver iterations:                             0


                 Variable           Value        Reduced Cost
                    P( 1)        65.00000            0.000000
                    P( 2)        65.50000            0.000000
                    P( 3)        66.00000            0.000000
                    P( 4)        124.3375            0.000000
                    P( 5)        66.00000            0.000000
                    P( 6)        66.00000            0.000000
                    P( 7)        80.00000            0.000000
                    P( 8)        65.50000            0.000000
                    P( 9)        66.50000            0.000000
                   P( 10)        65.50000            0.000000
```

图 6 局部最优解

根据上图，我们可以知道，第一类的局部最优解为 5070。也就是说，第一类的任务数为 78，会员数为 463 的情况下，我们求得最优解 5070，即第一类中的任务发布的总共的成本为 5070 元，平均成本为 65 元。之后的 15 个类别，也可以像上述的步骤进行求出局部最优解。

5.3.5 任务发布规则

根据 4.5，结合会员竞争力函数，给出发布任务的规则：1）若$x_{ijk}=1$，则在第一时间向第 k 个会员发布第 i 个区域中第 j 个任务的任务信息；2）建立任务—会员竞争力模型，即对第 j 项任务而言，会员的竞争力函数为：

$$y_{jk}=\frac{r_k}{d_{jk}}$$

在上述的模型中，我们可以由y_{jk}的值的大小顺序，向第 k 个人发布任务点的信息。

任务信息发布方案：

1）利用会员的竞争力模型排序，计算出会员竞争力的数值，然后对数值进行排序；

2）排好序后，按照排好的序号提取排名在前 50 的会员；

3）对这 50 名会员做分割，依次分割成 10 个小组，每个小组 5 人；

4）将打包好的任务首先分配给最优解，如果该最优解拒绝接受任务，则再将任务分配给会员竞争力排名第一的那个小组，如果改组会员仍然拒绝，则再将任务分配给下一排名的小组……直到打包好的任务被接受。

5.3.6 任务发布规划模型的评价

该模型充分利用 0-1 规划，实现了任务点与会员的最优分配，一方面节约了企业的成本；另一方面，让会员能在最短的时间内了解到任务的分配信息；再者，大幅度提高了任务被完成的几率。

5.4 问题四

5.4.1 论证不用重新聚类

对于问题四，给出了附录三新项目任务位置信息，如下图 7。首先将经度纬度直角坐标化，接下来我们是不需要重新设置聚类中心进行聚类的，其中原因有二：其一，根据其位置坐标作图，我们可以直观地看出，所有任务点都大致明显地分为了三类，然后对照问题一设置的聚类中心坐标进行比较发现，恰好有三个聚类中心分别落在了新项目任务的三簇中，正好是繁华商务中心或者交通枢纽处；其二，任务的定价及会员完成任务的情况都明显受地理位置直接或间接的影响，可以说是其任务定价的主要依据，因此在对新项目任务制定定价方案时，同样应该重视地理位置的影响，应该直接根据其与市中心（聚类中心）的距离进行分类并打包定价分发任务。

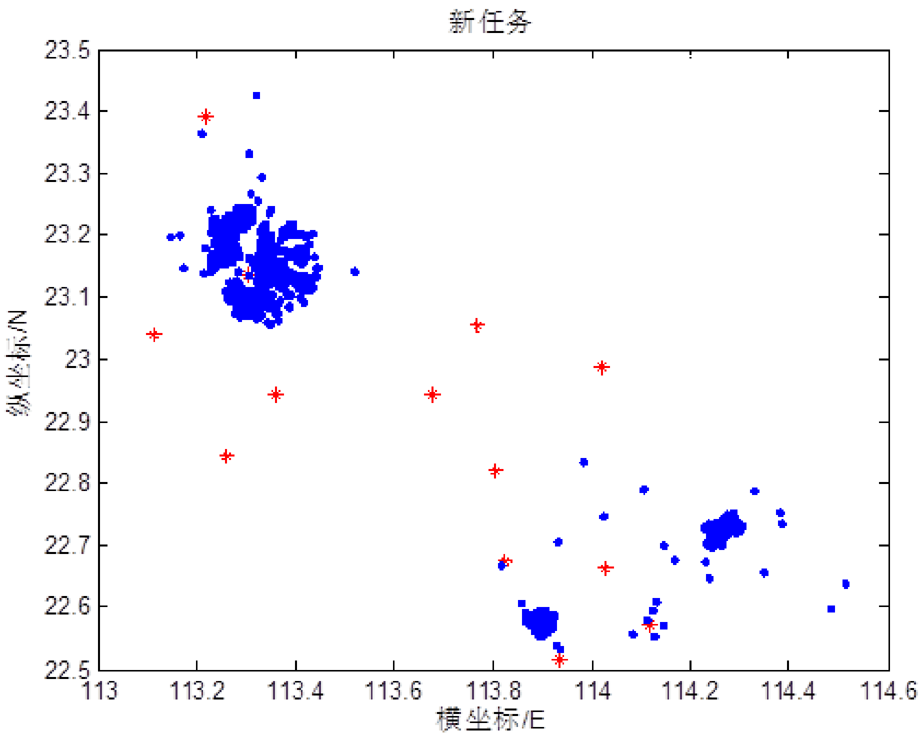

图 7 附件 3 的任务点分布图

由图 7 可知，新的任务点的分布主要是集中在 3 个大的片区，我们可以相信这些任务点主要都在这 3 个片区集群分布，所以我们不用再重新聚类了。

5.4.2 按上述模型带入求解

问题四要求我们对附件三中的新项目给出新的任务定价方案。由上面 5.4.1 节证明了我们可以应用问题一找到的聚类中心，将新项目任务应用 K 均值聚类分析法进行聚类。

（1） 聚类：

我们应用 K 均值聚类的方法，将附件三种任务点分为三类，同时可以得到每个任务点到其聚类中心的距离。

（2） 制定新任务点定价方案：

类似于问题二，利用 MATLAB 编程（程序代码见附录）求出某区域内任务点半径阈值为 K 的圆面内会员的个数，由于两影响因素的单位不同，需要分别将其单位化，然后代入问题二中所建立的任务点初步定价模型中，可以得到各任务点的回归价格，如下表 6 所示：

表 6 新任务点定价局部明细表

任务 GPS 经度	纵坐标	横坐标	CLUSTER	DISTANCE	distance'	n	n'	p
114.2408	-52315.58	96411.82	1	16747.39	-1.036290	69	0.077610	68.39301
114.2866672	-52057.53447	101103.6893	1	21063.35874	0.123399229	590	-1.146605008	70.7981471
114.2575	-53612.42	98120.00	1	17707.43	-0.778331	65	-0.382874	69.4610
114.3819	-49656.85	110864.7	1	31012.79	2.796793	43	-2.909925	73.34342
114.2721	-52983.38	99619.55	1	19333.55	-0.341394	63	-0.641194	69.7923
114.2732	-53502.59	99728.60	1	19237.83	-0.36711	63	-0.629963	69.7841

根据问题二和问题三，我们可以沿用上面的问题给出的模型，对其类均值和完成度做检测，则结果如下表 7 所示。

表 7 各类的均值和完成度

类别	类均值	完成度
第一类	88.98365	85.9
第二类	84.62636	88.4
第三类	79.79263	83.73

由表 7 中得出的数据，我们可以知道各类的均值都比之前的均值高，同时完成度也相应提高了。由此可见，我们的模型建立还是不错的，模型的拟合情况还是比较可观的。

下面，我们给出多元线性定价模型的回归方程：

$$p_{1j} = -0.91624x_1 - 2.83258x_2 + 67.66336$$
$$p_{5j} = -0.20288x_1 - 2.17438x_2 + 70.43590$$
$$p_{10j} = 2.12595x_1 + 0x_2 + 71.24341$$

6. 模型的评价及优缺点分析

模型的优点：

（1）问题一的模型比较简单明了，首先应用了 K 均值聚类分析的方法，将所有任务点划分为 16 类，并对照地图发现其聚类中心所对应的地区基本上都处于比较繁华的商业地带，这对后面影响因素的提出作出了相应的检验。然后对任务点标价、距聚类中心的距离作相关性分析，并线性回归得到一个直观的定价总体规律，最后还具体对每一类区域的任务点定价规律进行了阐述。这对一般性的数据统计操作者有很大的启发。

（2）问题二中主要是一个定价的优化模型。在原方案的定价规律基础上，考虑到了会员信息，找到一个优化的定价方案，从而在费用不增的条件下提高了任务的完成度。

（3）问题三考虑会员利益最大化的心理趋向，把分散的任务以任务点间的距离与原先设定的阈值相较大小的标准，进行打包处理。并借鉴了“拼车模型”的定价规律，对任务（包括任务包）再重新定价，从而任务的完成度提高了不少。最后应用 0-1 规划模型，以个人任务量限额为约束条件，其价格之和最小化的目标来分发任务。过程、目标明晰合理，并能够应用软件编程实现。

（4）问题四对新任务点进行入类处理，将以上经优化所得的定价模型应用其中，任务完成情况比较可观。

模型的缺点：

（1）在本文能够定量分析的影响标价因素有限，因此，所建立回归模型的效果不是很好。在实际生活中，直接不能利用该模型对任务后期的完成情况进行比较准确的把握。

（2）聚类分析的过程中，有以下两个不足之处：其一，由于一些不可控的因素，可能导致聚类中心的位置找的不是特别准确，从而影响到后面各种基于这种聚类结果所建立的模型；其二，任务点聚类后再将会员位置以同聚类中心再次聚类，得到的是同心圆环，因此划分的区域不能完全重合，而且很有可能还会与其他类的区域相重合或相离而留出空隙区域，对定价模型的制定有相当大的影响。

（3）对任务点进行打包处理时，其阈值的确定主观性很强，得不到一致的标准，由此建立的定价模型也有待考证。

模型的改进：

（1）综合多方面考虑任务点标价的影响因素，且找出它与一些有限的影响因子之间的关系进行深入研究，能够更精确的把握各任务点的标价规律，从而建立出更好的定价模型。

（2）由于导致某个任务点的标价较低的因素有很多，尽可能地找出多方面的因素， 及周边任务完成度的变化趋势，从而更准确地找出聚类中心的点，优化分类结果，改善模型效果。

7. 模型推广

本文所建立的定价模型不仅仅应用于本文的“照相赚钱”的任务中，还可以广泛地应用于交通运输的领域。在此，我们不仅仅要考虑到航程与起始地与目的地位置还需要考虑到一些人为因素和自然因素。此种模型可以应用到汽车的最优行驶。

问题三中，在对任务点打包后，建立的用户竞争模型在经济市场中应用十分广泛，类比于两家竞争公司，对于自己的商品，总是争相提高产品的评价等级（如信誉评分等级），为公司在市场竞争上占得一席之地。

对于任务的分发过程中建立的 0-1 规划模型也推广到社会的其他领域，可用来解决资源的有限条件下的调度问题，从而实现资源的有效利用，实现经济利益的最大化。

参考文献

[1] 黎珍惜 黎家勋.基于经纬度快速计算两点间距离及测量误差，http://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFQ&dbname=CJFDHIS2&filename=DBCH201311074&v=MjQ1NDk5TE5ybzlDWUlSOGVYMUx1eFlTN0RoMVQzcVRyV00xRnJDVVJMMmZidWRxRkNqbFZyeklJUy9JWnJHNEg=.2017.09.15.

[2] 王学民.应用多元分析[M].上海：上海财经大学出版社，2014.153-177

[3]韩中庚.数学建模方法及其应用[M].北京：高等教育出版社，2009.172-173、203-206

[4] 何晓群,刘文卿.应用回归分析（第四版）[M].北京：中国人民大学出版社，2015.57-60、70-75

[5] 何晓群.多元统计分析[M].北京:中国人民大学出版社,2015.

[6] 夏坤庄.深入解析SAS:数据处理、分析优化与商业应用[M].北京：机械工业出版社，2015.

[7] 谷鸿秋.SAS编程演义[M].北京:清华大学出版社，2017.

[8] 何正风.MATLAB R2015b 神经网络技术[M].北京:清华大学出版社，2016.

[9] 肖枝洪.时间序列分析与SAS应用[M].武汉:武汉大学出版社，2012.

[10] 余胜威.优化算法案例分析与应用(进阶篇)[M].北京:清华大学出版社，2015.

[11] MATLAB神经网络43个案例分析[M].北京:北京航空航天大学出版社,2013.

附录

附录一：数据导入

==

```
/*SAS中拓展语言*/
options validvarname=any validmemname=extend;
run;

/*数据导入*/

/*已结束项目任务数据*/
PROC IMPORT OUT= B.'已结束项目任务数据'n
          DATAFILE= "G:\全国大学生数学建模大赛正式比赛\B题\题目的信息\
附件一：已结束项目任务数据.csv"
          DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;
/*会员信息数据*/
PROC IMPORT OUT= B.'会员信息数据'n
          DATAFILE= "G:\全国大学生数学建模大赛正式比赛\B题\题目的信息\
附件二：会员信息数据.csv"
          DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;
/*新项目任务数据*/
PROC IMPORT OUT= B.'新项目任务数据'n
          DATAFILE= "G:\全国大学生数学建模大赛正式比赛\B题\题目的信息\
附件三：新项目任务数据.csv"
          DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;
/*对导入的缺省值进行处理*/
proc sql;
create table B.新项目任务数据 as select * from B.新项目任务数据 where 任务GPS
纬度~=.;
run;
quit;
```

附录二：数据可视化

==

```
/*数据可视化*/

/*所有会员（1877个）的地理位置可视化（大概）*/
axis1  order=(106 to 117 by 0.5)
       label=(h=1 '会员位置经度')
       minor=(color=blue number=1);
axis2  order=(20 to 34 by 0.5)
       label=(h=1 '会员位置纬度')
       minor=(color=red number=1);

proc gplot data=B.会员信息数据;
    title1 '会员位置';
    footnote h=1.5  "数据来源：题目中给出 " j=r'2017/09/14';
    plot 会员位置纬度 * 会员位置经度 /overlay legend haxis=axis1 vaxis=axis2;
    run;
    quit;
goptions reset=all;

/*已经结束项目的地理位置（835个）*/
axis1  order=(112.6 to 114.6 by 0.2)
       label=(h=1 '任务gps经度')
       minor=(color=blue number=1);
axis2  order=(22.4 to 24.0 by 0.2)
       label=(h=1 '任务gps纬度')
       minor=(color=red number=5);
symbol1 value=dot   cv=red  ci=red line=1;
proc gplot data=B.已结束项目任务数据;
    title1 '已经结束的任务位置';
    footnote h=1.5  "数据来源：题目中给出 " j=r'2017/09/14';
    plot 任务gps纬度 * 任务gps经度 /overlay legend haxis=axis1 vaxis=axis2;
    run;
    quit;
goptions reset=all;

/*新项目任务数据*/
axis1  order=(113.1 to 114.6 by 0.1)
       label=(h=1 '任务GPS经度')
       minor=(color=blue number=1);
axis2  order=(22.5 to 23.5 by 0.1)
       label=(h=1 '任务GPS纬度')
       minor=(color=red number=1);

proc gplot data=B.新项目任务数据;
    title1 '新任务位置';
```

```
    footnote h=1.5  "数据来源：题目中给出 " j=r'2017/09/14';
    plot 任务GPS纬度 * 任务GPS经度 /overlay legend haxis=axis1 vaxis=axis2;
    run;
    quit;
goptions reset=all;
```

附录三：坐标系的转换

==

```
data B.已结束项目任务数据经纬度换算;
set B.已结束项目任务数据;
纵坐标=(任务gps纬度-23)*111319.5;
横坐标=(任务gps经度-113)*102469.9;
run;


data B.会员信息数据经纬度换算;
set B.会员信息数据;
纵坐标=(会员位置纬度-23)*111319.5;
横坐标=(会员位置经度-113)*102469.9;
run;


data B.新项目任务数据经纬度换算;
set B.新项目任务数据;
纵坐标=(任务GPS纬度-23)*111319.5;
横坐标=(任务GPS经度-113)*102469.9;
run;
```

附录四： 转换坐标系后的数据可视化

==

```
/*寻找最低价格*/
proc sort data=B.已结束项目任务数据经纬度换算 out=one;
by 任务标价;
run;
proc sql;
create table two as select * from one where 任务标价<=67;
quit;
/*绘制图形进行可视化*/
axis1  order=(-80000 to 30000 by 10000)
      label=(h=1 '横坐标/米')
      minor=(color=blue number=1);
axis2  order=(-40000 to 120000 by 20000)
```

```
        label=(h=1 '纵坐标/米')
        minor=(color=red number=5);

proc gplot data=two;
    title1 '已经结束的任务位置（总图）';
    footnote h=1.5  "数据来源：题目中给出 " j=r'2017/09/14';
    plot 横坐标 * 纵坐标 /overlay legend haxis=axis1 vaxis=axis2;
    run;
    quit;
goptions reset=all;

/*宏程序放大图形，其中hb为横坐标起始点，he为横坐标终止点，zb为纵坐标起始点，ze为纵坐标终止点*/

%macro fangda(hb,he,zb,ze);
axis1  order=(&hb to &he by 10000)
        label=(h=1 '横坐标/米')
        minor=(color=blue number=1);
axis2  order=(&zb to &ze by 20000)
        label=(h=1 '纵坐标/米')
        minor=(color=red number=5);

proc gplot data=two;
    title1 '局部放大的图形';
    footnote h=1.5  "数据来源：题目中给出 " j=r'2017/09/14';
    plot 横坐标 * 纵坐标 /overlay legend haxis=axis1 vaxis=axis2;
    run;
    quit;
goptions reset=all;
%mend;
/*第一次放大*/
%fangda(-80000,-30000,-40000,20000);
%fangda(-80000,-30000,20000,110000);
%fangda(-30000,30000,-40000,20000);
%fangda(-30000,30000,20000,110000);
run;
```

附录五：聚类中心的选取

```
=========================================================================
/*选取20个聚类中心*/
proc fastclus data=two  maxc=20 radius=0 maxiter=10 out=聚类中心1;
var 横坐标 纵坐标;
```

```
run;

/*确定聚类中心*/
data center;
input 横坐标 纵坐标;
cards;
65314.09884 -73067.03509
77471.62223 -58703.81666
-2131.81699 3528.11718
26024.00383 -43709.58164
97142.29614 -56339.78565
2164.04850 -14400.16573
-4178.94386 -39665.86718
48274.60441 -17204.59051
4115.74452 -5398.30062
81894.40951 -70647.63177
-18898.72738 -18421.58026
56748.44672 -51950.41756
15817.06751 -8809.24362
-5834.86026 -6827.46643
8347.53762 -27094.15989
46469.17947 -43231.01428
55543.66356 -61721.89753
72904.39296 -24043.64945
-5705.75288 19755.10727
36790.79582 -26803.29898
run;
proc fastclus data=one seed=center maxc=100 radius=0 maxiter=10 out=聚类
中心2;
var 横坐标 纵坐标;
run;
/*对数据集进行排序*/
proc sort data=聚类中心2;
by CLUSTER DISTANCE;
run;
data 聚类中心2;
set 聚类中心2;
DISTANCE=DISTANCE/1000;
run;

ods graphics/ reset =all;
%macro jie(n);
proc sql;
create table leixing&n as select * from 聚类中心2 where CLUSTER=&n;
```

```
quit;

proc sort data=leixing&n;
by 任务标价;
run;

proc corr data=leixing&n  nosimple ;
var 任务标价 DISTANCE;
run;
%mend;
%jie(1);
%jie(2);
%jie(3);
%jie(4);
%jie(5);
%jie(6);
%jie(7);
%jie(8);
%jie(9);
%jie(10);
%jie(11);
%jie(12);
%jie(13);
%jie(14);
%jie(15);
%jie(16);
%jie(17);
%jie(18);
%jie(19);
%jie(20);
run;
```

附录六：线性回归

==

$$p_{1j} = -0.91624x_1 - 2.83258x_2 + 67.66336$$
$$p_{2j} = 0.64639x_1 + 0.36061x_2 + 66.47728$$
$$p_{3j} = 2.99377x_1 + 2.32364x_2 + 74.94559$$
$$p_{4j} = 1.61784x_1 + 0.88829x_2 + 72.10521$$
$$p_{5j} = -0.20288x_1 - 2.17438x_2 + 70.43590$$
$$p_{6j} = 0.21542x_1 + 0.94122x_2 + 72.26086$$
$$p_{7j} = 1.12581x_1 + 0x_2 + 70.19522$$
$$p_{8j} = -0.00928x_1 - 1.02650x_2 + 68.09783$$

$$p_{9j} = 0.34995\mathrm{x}_1 + 0.97722\mathrm{x}_2 + 66.01643$$
$$p_{10j} = 2.12595\mathrm{x}_1 + 0\mathrm{x}_2 + 71.24341$$
$$p_{11j} = 0.89004\mathrm{x}_1 + 0\mathrm{x}_2 + 68052759$$
$$p_{12j} = 1.16172\mathrm{x}_1 - 0.77783\mathrm{x}_2 + 70.82810$$
$$p_{13j} = 1.41164\mathrm{x}_1 + 0\mathrm{x}_2 + 69.01287$$
$$p_{14j} = 0.39552\mathrm{x}_1 + 1.10611\mathrm{x}_2 + 67.10375$$
$$p_{15j} = -2.26310\mathrm{x}_1 - 4.84255\mathrm{x}_2 + 68.87478$$
$$p_{16j} = -0.86639\mathrm{x}_1 + 0\mathrm{x}_2 + 69.40002$$