# “拍照赚钱”的任务定价

## 摘　要

“拍照赚钱”是众包中的一个平台，任务定价需综合考虑。首先通过对已结束项目数据分析，初步找出定价基本规律，考虑任务未完成的原因。影响定价的因素众多并且数据多，故才用 SPSS 来研究数据，进行 K 均值分类。对分类后的数据进行多元拟合分析，并完成任务定价。

针对问题一：初步采用多元拟合的方法对数据进行拟合分析，并进行相关性检验，发现相关性不高。由于数据量比较大，采用 SPSS 对数据进行 K 均值聚类，对聚类后的数据进行多元拟合，相关系数达到了 0.86625，聚类时发现任务定价同一类中在某个数值浮动。结合附件二中相关信息，任务未完成主要有以下可能性：会员信誉问题、任务所在地的经济问题、会员能力问题以及任务分配不合理。

针对问题二：采用模型一的解题思维，并对所有数据进行综合考虑，K 均值聚类时，注重信誉的问题，同时需要将任务位置和会员位置聚类后用最短距离法相关联，多元拟合出任务定价的回归方程。通过该方程可以求出任务聚类中心处的定价，利用任务位置与每个聚类中心的距离关系进行逐个定价，该模型比模型一相关性高，且定价合理，同时有助于提高任务的完成度。

针对问题三 本问题需要把一些位置比较集中的任务点联合在一起进行打包发布。通过聚类分析分出一定类别且容量不同的类，首先需要人为设置一个阈值筛选出适合进行打包发布的数据，以一个类的任务中心点代替那些适合进行打包发布的任务点与不适合进行打包处理的任务点进行数据拟合。在聚类分析和数据拟合过程中需要考虑附件二中会员的信誉值和预定限额。其次还从多个方面分析了打包后对任务完成情况的影响。

针对问题四 本问题通过在前三个问题所建立的模型分别建立三个不同的任务定价方案，通过在相关系数 R，F 检验，模型考虑的方面已及任务的完成度等方面进行横向对比。分析出模型三在相关系数，F 检验值上有突出的优点，完成率最高。

通过对四个问题的分析与解答，“拍照赚钱”这种互联网的众包模式运用 SPSS 软件对影响任务定价的因素：任务地点、会员地点、信誉值、预定限额以及打包处理进行了逐层逐类的分析，可以综合考虑得出使成本最小化的数学模型，以供这类行业者进行参考。

**关键字**：SPSS；K 均值聚类；多元拟合；最短距离法

# 一、问题重述

## 1.1 问题背景

近年来，随着互联网的兴起与发展，众包[1]模式也在悄悄地发展，改变我们生活中的商业模式。众包：一个公司或机构把过去由员工执行的工作任务，以自由自愿的形式外包给非特定的大众网络的做法。“拍照赚钱”是一个典型的众包式自助服务 APP，这种基于互联网的自助式服务平台，不仅可以有效的为企业提供商业信息；相比传统的市场调查模式，还可以大大的减少任务完成成本。APP 是众包服务平台的核心，任务的定价是任务完成效率的核心因素，如何制定一个完成率高、合理的定价方案有待研究。

## 1.2 问题相关信息

根据对“拍照赚钱”APP 的相关信息的研究，给出了以下 3 个附件相关数据。

附件一：一个已经结束的项目任务 835 组数据，包含了每个任务的位置、定价和任务完成情况。

附件二：会员的信息表，包含了会员的位置、誉值和根据信誉值参考得到的预定任务限额和预定任务开始的时间(任务的分配跟预定额所占比例进行分配)。

附件三：一个新的检查项目任务数据，只有任务位置。

## 1.3 问题的提出

根据题中所给以上信息，本文将定价方案细化为以下四个问题，并建立数学模型进行分析、比较和改进。

问题一：找出附件一中项目的任务定价规律，分析附件一中任务失败的原因。

问题二：为附件一中的项目任务设计新的定价方案，并与原方案进行对比。

问题三：在一些实际情况下，有些任务因为位置的集中，会引起会员的争夺，一种考虑是将这些任务联合打包发布。在这样的考虑下，怎么修改前面的定价方案，对任务最后的完成又有什么影响？

问题四：对附件三中给出的新项目制定定价的方案，并评价方案对任务的完成效果。

# 二、字符说明

表 2-1 字符说明

| 字符 | 字符意义 |
|---|---|
| D | 任务定价 |
| P | 任务完成情况 |
| $X_1$ | 任务位置纬度 |
| $X_2$ | 任务位置经度 |
| $X_3$ | 会员位置纬度 |
| $X_4$ | 会员位置经度 |

| | |
|---|---|
| $\varphi(x)$ | 会员信誉 |
| $d_{ij}$ | i 和 j 之间的距离 |
| $D_j$ | j 类聚集中心的定价 |
| $D_i$ | 第 i 个任务的定价 |
| R | 相关系数 |

# 三、问题假设

（1）拍钱赚 APP 是静态的，即工作者的数量不随时间的推移而产生改变；

（2）每一个工作者都是理性的，期望最大化收益；

（3）忽略第三方软件的影响；

（4）任务一旦分配则不能改变；

（5）会员完成任务的意志与获得酬劳的动机和获得认可的动机成正相关；

（6）每一个任务都是独立的且不相互影响。

# 四、问题分析

## 4.1、问题一分析

该问主要是分析并给出任务未完成的原因，通过对附件一的数据分析，得到的有效信息主要是每个任务的位置、定价和完成情况。初步决定对附件一中任务的经纬度和定价建立多元拟合，来分析任务位置和任务定价的基本规律，接着我们还可以结合附录二中工作者的位置的数据，与附录一中有效的信息进行多元拟合，建立任务定价关于位置的模型，接着进行相关系数检验，如果相关系比较满足要求，就以此为模型，否则利用聚类分析[2]的方法进行任务进行分类，在对这 K 个类别进行多元拟合，通过对定价规律的分析，来研究任务未完成的原因。

## 4.2、问题二分析

问题二的求解，主要是在问题一的任务未完成的原因上建立新的模型，对任务进行重新定价，因为问题一的模型是对任务位置进行建立的模型，没有进行多方面考虑。问题二将会员的信誉、位置和任务位置的进行聚类分析求解出新的定价方案。

## 4.3、问题三分析

问题三是在任务位置比较集中的条件下，考虑将集中的任务进行打包发布，对问题二中模型进行点化的聚类分析，提出不同任务离聚类分析中点距离的分析，建立打包发布的中心定价与会员位置和任务位置等的数学模型，并对任务最终完成的影响进行分析。

## 4.4、问题四分析

问题四是对附件三中的新项目数据制定一个定价方案，首先对附件三中的

项目任务进行地图定位分析，分析可得，任务的位置大体呈现出聚集的现象。利用问题三的数学模型制定定价方案，研究该方案实施效果的。

# 五、模型的建立与求解

“拍照赚钱”APP 实际就是众包领域中的一个小平台。所谓众包[2]是随着互联网兴起，公司等机构寻求从互联网上解决问题途径的一种新型的商业模式。“拍照赚钱 APP”的工作流程和众包的工作流程相似，如下图 5-1 所示。

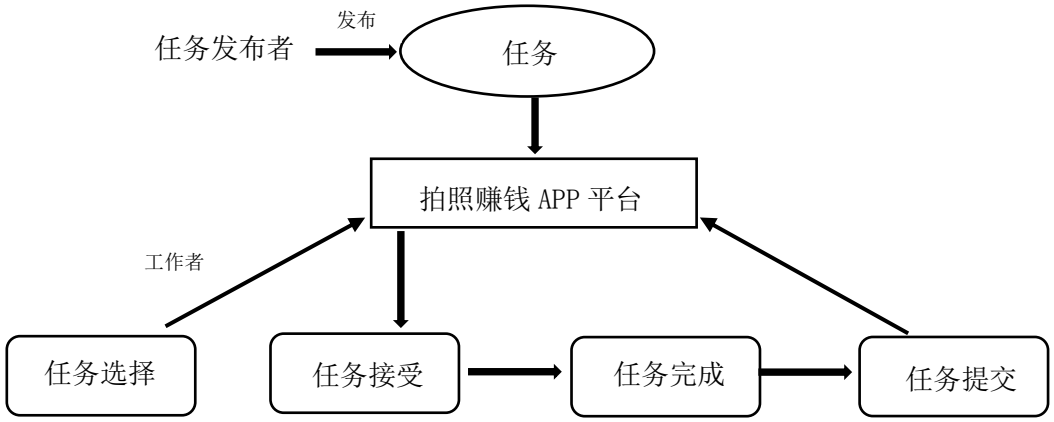


图 5-1“拍照赚钱 APP”的工作流程

## 5.1、问题一模型的建立与求解

### 5.1.1、模型的建立

问题一中需要解决的问题有两个，一个是项目的定价规律，另外一个就是利用上述求得的定价规律来分析附件一中任务未完成的原因。附件一中的数据只有五项，即数据号码、任务纬度wd、任务经度jd、任务标价D以及任务执行情况p如下式（5-1）所示。

$$p=\begin{cases}0, & (\text{未完成任务})\\ 1, & (\text{完成任务})\end{cases} \qquad (5\text{-}1)$$

经过对附件一中数据的分析，对任务标价的有效信息为任务纬度wd、任务经度jd，对于任务点的分布情况，可以通过 MATLAB 绘图来表示，具体的分布情况如下图 5-2 所示。

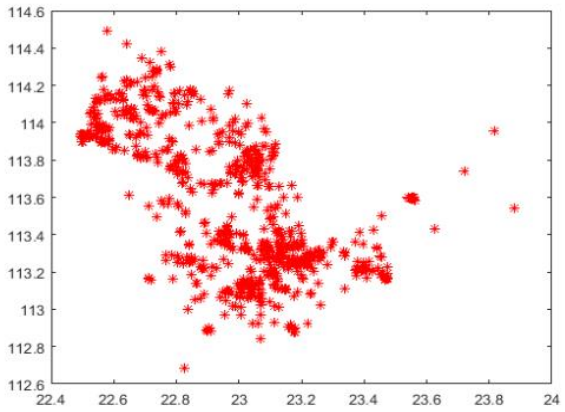


图 5-2 任务点的位置分布图

由上述对附件一中数据的分析，初步建立定价关于任务点所在位置及经纬度的模型，主要是运用多元插值拟合的数学思想，其中的多元就是任务点的纬度和经度的坐标，此处分别用 $x_1$ 和 $x_2$ 来表示。在多元拟合过程中，分别进行一次多元拟合如下式（5-2）和二次多元拟合如下式（5-3），并进行做图对比分析如图 5-3 所示，得到的结果如下

$$D = -19.437 + 2.4678x_1 + 0.2785x_2 \quad (5\text{-}2)$$

$$D = 255270 - 2470.7x_1 - 3996.2x_2 + 6.1431{x_1}^2 + 19.301x_1x_2 + 15.646{x_2}^2 \quad (5\text{-}3)$$

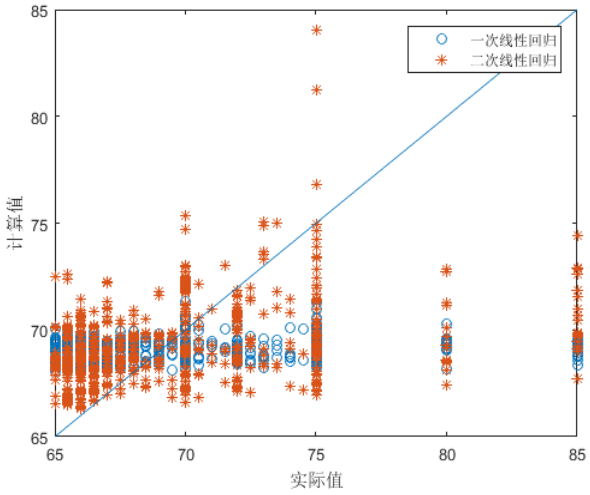


图 5-3 拟合真实值和计算值对比

从图 5-3 中我们可以看出一次线性回归比二次线性回归相对而言好一点，二次线性回归的数值振动幅度比较大，然而一次线性回归更趋于稳定，故采用一次线性多元拟合回归做为初步模型即：

$$D = -19.437 + 2.4678x_1 + 0.2785x_2 \quad (5\text{-}4)$$

经过对附件二中的数据进行简单的分析，在结合建立的初步模型，发现如果仅仅利用任务的位置来进行多元拟合的话，比较单一，误差较大，偶然性也比较大。由于附件一中的数据是已经结束的项目，在进行的时候对于会员的信誉度[3]了解比较少，但是对于会员所处的位置是知道的，此时我们建立关于任务的位置

和会员的位置的多元拟合，其中 $x_3$ 和 $x_4$ 分别表示会员所在地的纬度和经度坐标，拟合之前需要对会员所在地的数据进行筛选[6]，然后再利用附录中的多元拟合程序得到以下拟合函数：

一次拟合：

$$D = -63.9312 + 2.4646x_1 + 0.3164x_2 - 0.0961x_3 + 0.375x_4 \quad (5\text{-}5)$$

二次拟合：

$$D = -282830 - 2889.4x_1 - 4158.2x_2 - 61.3064x_3 - 226.8802x_4 + 6.6291{x_1}^2 + 19.274x_1x_2 + 1.0713x_1x_3 + 3.2983x_1x_4 + 15.7105{x_2}^2 + 0.2697x_2x_3 + 1.2470x_2x_4 - 0.0803{x_3}^2 + 0.0924x_3x_4 + 0.0355{x_4}^2 \quad (5\text{-}6)$$

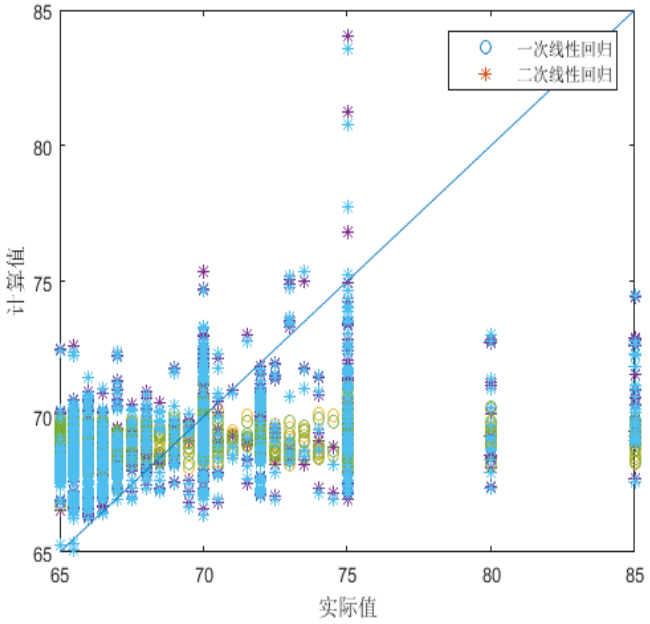


图 5-4　拟合真实值和计算值对比

从图 5-4 中我们可以得出一次多元拟合比较好，实际值和拟合计算值得相差比较小即计算误差小。综合初次模型通过多元拟合得到任务定价模型为：

$$D = -63.9312 + 2.4646x_1 + 0.3164x_2 - 0.0961x_3 + 0.375x_4 \quad (5\text{-}7)$$

在对模型进行相关系数分析的时候，我们发现模型的相关系数为 0.15，可信度不高，此处可以得到对于数据不能直接就进行拟合，需要做些处理。故我们决定先对对数据进行聚类分析，然后再进行多元拟合。

聚类分析就是分析如何对样品和变量进行量化分类，此处我们采用 K 均值聚类分析。K 均值聚类是以距离的远近亲疏为标准进行聚类的，K 均值聚类可以产生指定类数的聚类结果。由于附件一中的数据较多我们将数据一共分为 10 类，得到结果如下表所示

表 5-1 K 均值聚类结果

| 序号 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 纬度 | 23.06 | 22.75 | 22.99 | 23.04 | 22.58 |
| 经度 | 113.11 | 114.38 | 113.02 | 112.92 | 114.16 |

| | | | | | |
|---|---|---|---|---|---|
| 价格 | 66.50 | 70.00 | 69.00 | 71.50 | 68.00 |
| 序号 | 6 | 7 | 8 | 9 | 10 |
| 纬度 | 22.58 | 22.66 | 22.78 | 23.01 | 23.18 |
| 经度 | 114.49 | 114.07 | 114.30 | 112.97 | 112.88 |
| 价格 | 85.00 | 65.00 | 73.00 | 75.00 | 80.00 |

表 5-1 为利用聚类分析后得到的 10 组类别的中心坐标，对于附件一中一共用 835 组数据，每一类中所包含的具体数如下表 5-2 所示

**表 5-2 每一类中所包含数据个数**

| | | |
|---|---|---|
| 聚集 | 1 | 192.000 |
| | 2 | 65.000 |
| | 3 | 70.000 |
| | 4 | 77.000 |
| | 5 | 102.000 |
| | 6 | 27.000 |
| | 7 | 189.000 |
| | 8 | 20.000 |
| | 9 | 80.000 |
| | 10 | 13.000 |
| 有效 | | 835.000 |

从上表我们可以得出每一类中数据越多，定价的范围越大，同时在同一类中数据里该类中心的距离越远的话，其定价就会越大，离中心越近，该任务的定价就相对较少，从表 5-3 可以看出（由于类别中的数据比较多故选取其中一类进行分析）。

我们对以上 10 类数据进行多元拟合。

**表 5-3 同一聚类中价格变化**

| 任务 | 类 | 定价 | 任务 | 类 | 定价 |
|---|---|---|---|---|---|
| 443 | 6 | 75 | 496 | 6 | 85 |
| 446 | 6 | 67 | 497 | 6 | 65.5 |
| 453 | 6 | 67 | 499 | 6 | 75 |
| 454 | 6 | 70 | 500 | 6 | 67 |
| 455 | 6 | 70 | 501 | 6 | 67 |
| 464 | 6 | 73 | 503 | 6 | 70 |
| 469 | 6 | 75 | 511 | 6 | 70 |
| 477 | 6 | 73 | 512 | 6 | 73 |
| 479 | 6 | 67 | 518 | 6 | 75 |
| 481 | 6 | 70 | 520 | 6 | 73 |
| 492 | 6 | 65.5 | 522 | 6 | 67 |
| 493 | 6 | 65 | 755 | 6 | 70 |
| 494 | 6 | 69 | 765 | 6 | 65.5 |

拟合分析如得到的模型为：

$$D = 95111 + 8922.3x_1 - 3467.9x_2 - 43.687x_1^2 - 60.943x_1x_2 + 21.39x_2^2 \quad (5\text{-}8)$$

真实计算值和实际值如下图 5-5 所示：

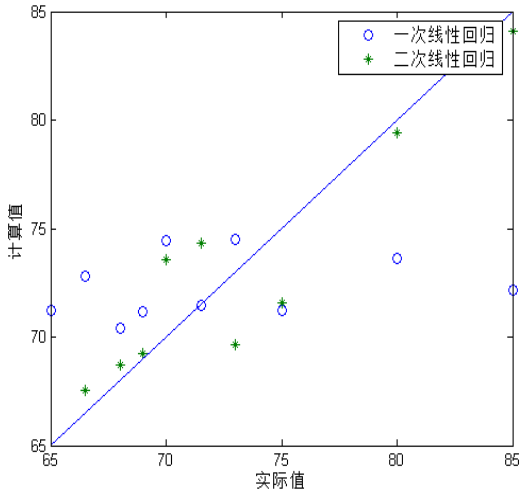


图 5-5 计算值和实际值

通过 MATLAB 分析对上述所用到的模型进行相关性检验，相关性 R=0.86625，可以看出该拟合效果较好，同时 F 值为 5.1815，P 值为 0.068061，小于 1，整个模型的拟合效果满足要求。

### 5.1.2、原因分析

结合过对定价规律以及下述分析得到以下原因：

1、会员的信誉度与完成任务相关，未完成区域中信誉较差的人员较多；

2、任务的分配不合理。未完成区域中信誉高和信誉低的人有较大的交叉，且信誉低的会员数量远远高于信誉高的人数；

3、会员能力有差异，信誉低的能力普遍比信誉较高的能力差；

4、未完成任务部分位置处于偏远；

5、经济因素的影响。

对以上原因进行具体分析针对附件一中任务未完成的原因，首先我们需要画出附件一中任务完成和未完成的任务点地理位置分布，如下图 5-5 所示。

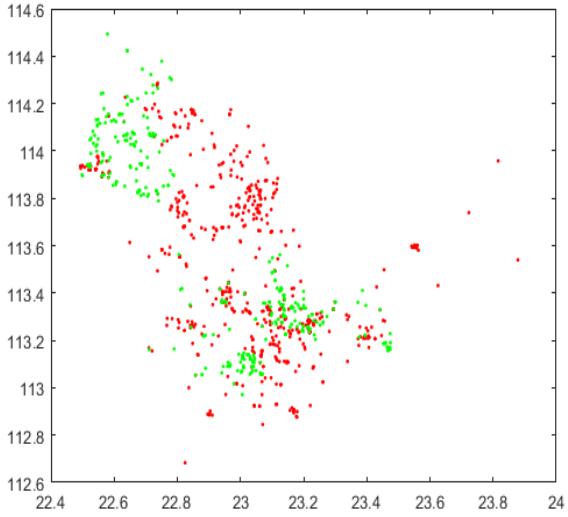


图 5-5 任务完成情况分布图

图 5-5 中青色的圆点表示任务未完成的位置分布，红色的圆点表示任务完成的位置分布情况，同时坐标是以任务位置的纬度为 x 轴，经度为 y 轴。从图中我们看到完成情况和未完成情况的分布比较集中，此时我们预估可能与会员所在的位置即会员在任务所在地的人数有关，这就需要我们作出任务完成情况以及会员位置的分布图，如下图 5-6 所示。

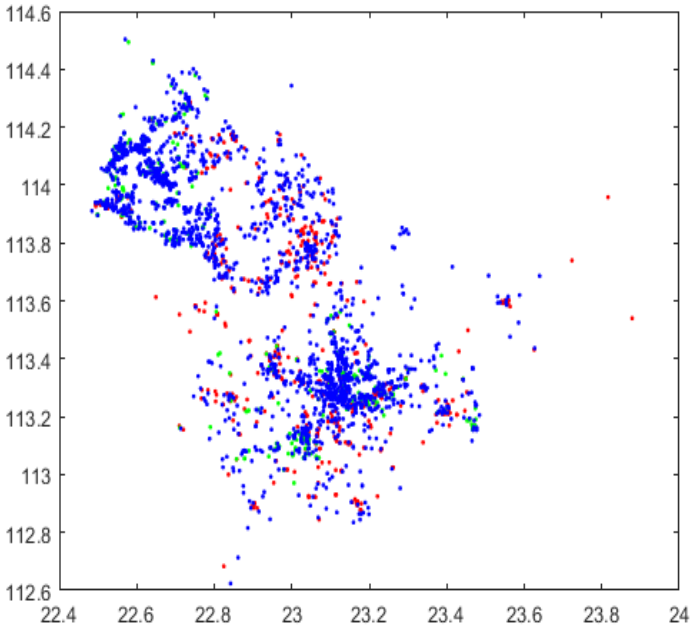


图 5-6 任务完成情况及会员分布图

图 5-6 中蓝色的点表示会员的分布情况，其他颜色的点表示意义和图 5-5 中相同，此处不再说明。从图 5-6 中可以看出会员比较集中的地方任务完成的情况都比较少，这说明这任务完成的情况和会员的数量没有较大的关系，为了更好的来解决任务为什么没有完成，我们在图中标出会员信誉分布情况如图 5-7 所示，

同时我们用下式来评价会员信誉度$\varphi(x)$的高低

$$\varphi(x)=\begin{cases} x>1000 & \text{（信誉度高）} \\ 100<x\leq 1000 & \text{（信誉度较高）} \\ 20<x\leq 100 & \text{（信誉度中）} \\ 10<x\leq 20 & \text{（信誉度低）} \\ x\leq 10 & \text{（信誉度差）} \end{cases} \tag{5-9}$$

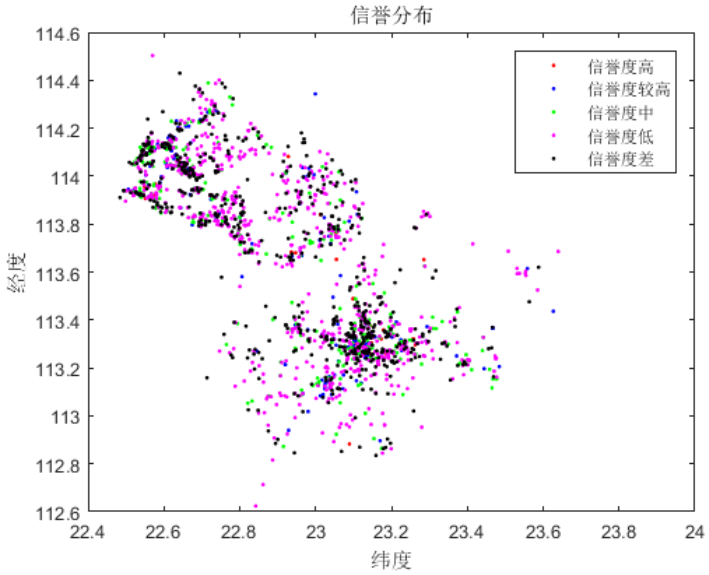


**图 5-7　会员信誉分布情况**

从图 5-7 中可以看出会员信誉低和差的人数较多，同时信誉好的比较少，这对于任务的完成有一定的影响，我们也可以看到信誉高的人和信誉差的会员有大量的交叉，这对于任务的分配可能会出现过分配不均。同时再结合图 5-5 中任务完成情况，我们可以看到在任务为完成区域，信誉差和低的会员数量较多，初步我们可以得到任务为完成可能与会员的信誉有关。对附件二中的会员预定的任务数中我们可以得到每个人的能力也是不同的，一般信誉较高的会员，其预定的任务数量也就越多，侧面也能折射出能力的差异。

经过上述的分析我们得到以下几点为完成任务的原因：

1、会员的信誉度与完成任务相关，未完成区域中信誉较差的人员较多；

2、任务的分配不合理，未完成区域中信誉高和信誉低的人有较大的交叉，且信誉低的会员数量远远高于信誉高的人数；

3、会员能力有差异，信誉低的能力普遍比信誉较高的能力差；

4、未完成任务部分位置处于偏远。

5、经济因素的影响。

## 5.2、问题二模型的建立与求解

### 5.2.1、模型的建立

经过对问题二的分析，初步得出以下几点：

(1)任务定价与任务的地理位置有关以及任务的完成有关；

(2)任务的定价与会员的地理位置有关；

(3)任务的定价可能与会员的信誉度值、开始预订的时间以及预订限额有关。

在分析中我们发现会员开始预订时间和预订的限额均是参考会员自身的信誉值，从这可以体现信誉值是十分重要的，尤其是在对任务的定价过程中有着十分重要的参考价值。如果将任务定价视为是关于以上分析中各因素的函数的话，定于会员的信誉度有关。

为了更好的给附件一中的任务定价，在建立模型的过程中参考问题一中的部分求解思路。我们利用 K 均值聚类分析的方法来处理数据，主要是由于附件数据比较多，通过 MATLAB 做散点图，发现任务分布没有明显的规律。附件中数据还给出了任务的所在地以及会员所在位置的情况，利用距离之间的关系来得出位置和价格的规律。首先利用 SPSS 对附件一中的数据进行分类[5]，其中变量分别为任务所在地的经纬度、任务的定价还有完成情况，对附件中 835 组数据分类成 16 大类，每一类中均有不同的任务，我们见其中的任务视为其元素，具体分类情况如下表 5-4 所示.

表 5-4 任务类的聚集中心

| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 纬度 | 22.64 | 23.10 | 22.72 | 22.58 | 22.82 | 22.57 | 22.54 | 22.84 |
| 经度 | 114.42 | 113.75 | 114.32 | 114.49 | 112.68 | 114.18 | 114.07 | 113.00 |
| 任务标价 | 75.00 | 71.00 | 70.00 | 85.00 | 75.00 | 73.00 | 80.00 | 73.00 |
| 序号 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 纬度 | 23.06 | 22.66 | 23.20 | 22.99 | 22.72 | 23.06 | 23.47 | 23.18 |
| 经度 | 113.11 | 114.07 | 113.17 | 113.07 | 114.28 | 112.92 | 113.16 | 112.88 |
| 任务标价 | 66.50 | 65.00 | 69.50 | 68.00 | 67.00 | 85.00 | 72.00 | 80.00 |

各类中元素数目（及任务的个数）如下表 5-5 所示。

表 5-5 每一类聚集观察值个数

| | | | |
|---|---|---|---|
| 1 | 37.000 | 9 | 190.000 |
| 2 | 49.000 | 10 | 133.000 |
| 3 | 63.000 | 11 | 72.000 |
| 4 | 11.000 | 12 | 79.000 |
| 5 | 44.000 | 13 | 81.000 |
| 6 | 9.000 | 14 | 16.000 |
| 7 | 4.000 | 15 | 22.000 |
| 8 | 16.000 | 16 | 9.000 |

对于附件二中会员的地理位置进行分类，经过对附件数据的分析和以上问题的求解我们可以得出会员的信誉值对于任务的定价有着很大的影响，同时还与会员预订限额和预定时间有关，我们分类是主要是通过信誉来进行分类，分 16 类其结果如下表 5-6 所示：

表 5-6 会员位置分类聚集中心

| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 经度 | 22.95 | 22.58 | 23.19 | 23.26 | 33.65 | 22.26 | 29.56 | 23.14 |
| 纬度 | 113.68 | 113.97 | 113.35 | 113.32 | 116.97 | 112.80 | 106.24 | 113.38 |
| 预定任务限额 | 114.00 | 163.00 | 139.00 | 98.00 | 66.00 | 72.00 | 15.00 | 95.00 |
| 序号 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 经度 | 23.19 | 23.15 | 22.87 | 22.80 | 23.10 | 23.70 | 22.97 | 22.99 |
| 纬度 | 113.57 | 113.31 | 113.65 | 113.36 | 113.63 | 113.11 | 113.60 | 113.87 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 预定任务限额 | 87.00 | 85.00 | 95.50 | 41.00 | 179.67 | 63.00 | 5.17 | 111.50 |

对任务和会员分类完成后我们需要建立起会员和任务所在位置的关系，此处我们采用距离最小法来将会员中的类与任务分的类进行相关联，我们可以将两大类的聚集中心求距离来是最小的两类来进行关联.

$$d_{ij} = \sqrt{(xi - xj)^2 + (yi - yj)^2} \tag{5-10}$$

经过我们计算得到的关联结果如下表 5-7 所示：

**表 5-7 关联结果**

| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 定价 | 75 | 71 | 70 | 85 | 75 | 73 | 80 | 73 |
| 任务纬度 | 22.64 | 23.1 | 22.72 | 22.58 | 22.82 | 22.57 | 22.54 | 22.84 |
| 任务经度 | 114.42 | 113.75 | 114.32 | 114.49 | 112.68 | 114.18 | 114.07 | 113 |
| 会员纬度 | 22.58 | 23.05 | 22.95 | 23.29 | 33.65 | 23.14 | 23.19 | 23.13 |
| 会员经度 | 113.97 | 113.65 | 113.68 | 113.65 | 116.97 | 113.38 | 113.35 | 113.24 |
| 序号 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 定价 | 66.5 | 65 | 69.5 | 68 | 67 | 85 | 72 | 80 |
| 任务纬度 | 23.06 | 22.66 | 23.02 | 22.99 | 2.72 | 23.06 | 23.47 | 23.18 |
| 任务经度 | 113.11 | 114.07 | 113.17 | 113.07 | 114.28 | 112.92 | 113.16 | 112.88 |
| 会员纬度 | 23.15 | 23.18 | 23.22 | 27.12 | 23.26 | 22.26 | 22.26 | 29.56 |
| 会员经度 | 113.31 | 113.18 | 113.29 | 111.02 | 113.32 | 112.08 | 112.08 | 106.24 |

经过分析，我们决定对上述 16 组数据进行多元拟合的方法，求出分类的中心处坐标关于定价[6]的模型，然后我们在到单个类中，每一个元素与聚集中心坐标的距离关系，再建立起相应的定价方程。

聚类后进行一次拟合和二次拟合得到方程为：

$$\mathrm{D} = 717.4987 - 10.1194x_1 - 2.7657x_2 - 0.0679x_3 - 0.8561x_4 \tag{5-11}$$

$$\begin{aligned}\mathrm{D} = {} & 430770 - 5312x_1 - 2680.5x_2 - 9326.8x_3 - 1949.1x_4 + 91.9795x_1^2 \\ & +41.1771x_1x_2 + 99.8465x_1x_3 - 59.1692x_1x_4 - 15.4145x_2^2 + \\ & 56.9201x_2x_3 + 33.1496x_2x_4 + 3.1167x_3^2 + 3.9612x_3x_4 - 2.358x_4^2\end{aligned} \tag{5-12}$$

同时利用 MATLAB 画出计算值和实际值图像如下图 5-8 所示：

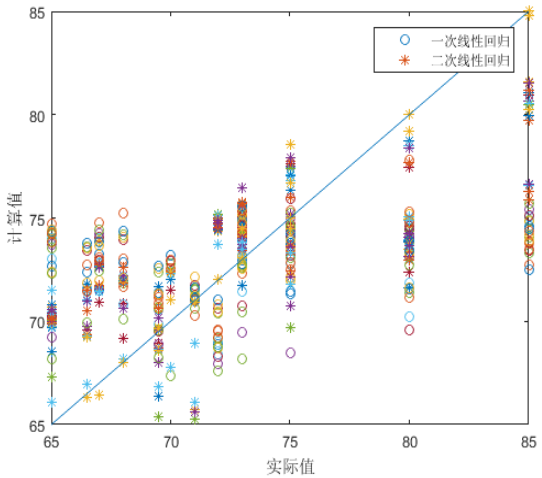


**图 5-8 计算值和实际值对比**

拟合是计算出一次拟合的相关系数为 0.1405,相关行比较低，二次拟合得到的相关系数 r=0.9920 相关性很好,我们采用多元拟合二次回归作为每一类中任务聚集中心价格求解的模型。

$$D = 430770 - 5312x_1 - 2680.5x_2 - 9326.8x_3 - 1949.1x_4 + 91.9795x_1^2 + 41.1771x_1x_2 + 99.8465x_1x_3 - 59.1692x_1x_4 - 15.4145x_2^2 + 56.9201x_2x_3 + 33.1496x_2x_4 + 3.1167x_3^2 + 3.9612x_3x_4 - 2.358x_4^2 \quad (5\text{-}13)$$

利用上述模型我们可以求出分类中的聚集中心处的任务定价，接着就是求出每一个任务的定价$D_{\mathrm{i}}$,在计算前我们已经对任务进行分类，且与任务的聚类进行了关联即我们已知每个任务所在的类中，我们可以通过将单一任务的地理位置坐标与该类聚集中心处的坐标求出距离来进行价格的敲定，同时我们需要考虑在该类任务中会员人数和任务数的关系来，因为任务定价一般与会员的密度成反比。

$$D_i = \begin{cases} D_j - q\mathrm{d}_{ij}, \text{ j 类中会员多于任务} \\ D_j + q\mathrm{d}_{ij}, \text{ j 类中会员少于任务} \end{cases} \quad (5\text{-}14)$$

$$\mathrm{d}_{ij=}\sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \quad (5\text{-}15)$$

式中：$D_{\mathrm{i}}$:单一任务的定价；

$D_{\mathrm{j}}$:每一类中聚类中心处的定价；

$x_{\mathrm{i}}, x_{\mathrm{i}}, x_{\mathrm{j}}, y_{\mathrm{j}}$:单个任务的地理位置坐标，和聚类中心处的坐标

$q$:第 j 中单位距离价格改变量；

j=1、2、3…16； i=1、2、3…835；

$\mathrm{d}_{ij}$：第 j 类中的 i 任务与其聚类中心的位置。

经过对 K 均值聚类分析，得到类型数据进行处理和分析得到 q 的范围一般在 10-15 以内。

### 5.2.2、与模型一比较分析

问题一中的模型建立主要是关于任务位置的经纬度，通过 K 均值聚类分析，通过将距离相对比较近的分到一组中，再利用多元拟合的方法来求解任务定价的原因。模型一考虑的内容比较少。该模型与模型二相比较，就相当简单，模型二的建立也是采用先聚类[9]在多元拟合的手段。但是模型二中进行分类时参考的变量比较全面，充分考虑影响任务定价的因素诸如任务的地理位置、会员的地理位置、会员预定限额、预定时间以及会员的信用度。这些不仅影响着任务的定价还会对任务的完成情况有这比较显著的影响。

模型一和模型二中的相关性检验也不与相同，模型一的相关系数 R=0.86625，其 F 检验中 F 值为 5.1815，相关性还行。模型二的相关系数 R=0.992，其 F 值为 8.24，和模型一相比，均得到了显著的提升，同时模型二在任务的定价更加的合理，从分考虑了会员密度对任务定价的影响，会员人数一般与定价成反比，模型二中分别对任务和会员分类，再利用最短距离法将任务雨会员关联避免任务未完成这种情况的出现，提高任务的完成率。

## 5.3、问题三模型的建立与求解

### 5.3.1、模型的建立

经过对问题三的深入分析，我们可以在问题二的基础上进行优化、分析进行求解。在位置集中处进行任务联合打包[8]处理时，我们假设打包 i 个任务，个体离任务中心点的距离可以用 di 来表示。我们运用 K 均值聚类分析的方法将任务点分为 p 类，设定能够进行联合打包的最低容量$S_{min}$，如果在第 j（j=1, 2……p）类中的容量大于$S_{min}$我们就将 j 类任务进行联合打包处理，反之，如果在第 j（j=1,2……p）类中的容量小于$S_{min}$我们就不能对其中的元素进行打包，只能单独的研究每一个任务点。我们将任务中适合打包的元素打包之后，选取打包元素的中心点来代替之后的数据处理过程。

我们在问题二的基础上可以求出在各类中心点以及未打包任务点的任务定价。在处理各类中元素与中心点的距离时我们运用已经建立的距离价格公式（$d_{ij}=\sqrt{(xi-xj)^2+(yi-yj)^2}$）进行对各个类中的元素定价。

### 5.3.2、模型的求解

我们对任务地点的经度和纬度运用 SPSS 软件进行 K 均值聚类分析。选取分的类数为 10 类，将各类元素大于 50（即$S_{min}$=50）的类看作可以进行联合打包处理。将第 1、3、4、5、8、9、10 类的中心点看作为一个任务点，其与 2、6、7 类中的元素组成一个容量为 64 的新任务点数组。

在聚集数目表中除了几个容量较小之外基本都在 100 容量以上，这说明分类的数目基本合理，分类的方法比较合适。

**表 5-8 聚集观测数目**

| 聚集中的观测数目 | | |
|---|---|---|
| 聚集 | 1 | 105 |
| | 2 | 17 |
| | 3 | 179 |
| | 4 | 134 |
| | 5 | 98 |
| | 6 | 37 |
| | 7 | 3 |
| | 8 | 101 |
| | 9 | 61 |
| | 10 | 100 |
| 有效 | | 835 |

**表 5-8 聚集中心**

| 序号 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 纬度 | 23.26 | 23.88 | 23.39 | 22.86 | 22.58 |
| 经度 | 223.02 | 113.54 | 113.41 | 113.14 | 114.49 |
| 序号 | 6 | 7 | 8 | 9 | 10 |
| 纬度 | 22.82 | 23.82 | 22.97 | 222.81 | 22.5 |
| 经度 | 112.68 | 113.96 | 113.99 | 113.56 | 113.9 |

在联合打包后画出任务地点分布图如下图 5-9，其分别范围和趋势大致和原任务地点分布图一致，说明联合打包后依然具有与原始数据的统一性和一致性，但是在数据处理的复杂程度上，打包后的数据处理具有明显的优势。

此时我们可以继续利用问题二的模型以各类中心价格为起点计算任务地点的定价。

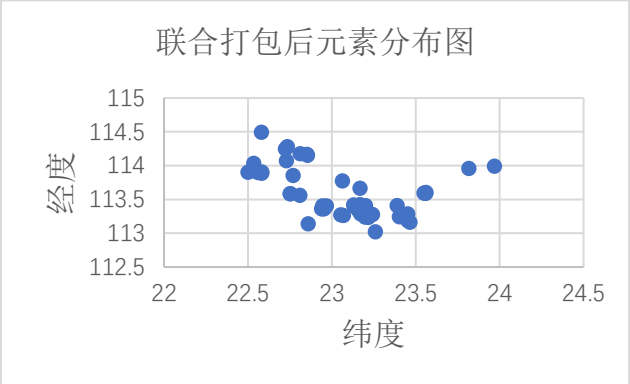


图 5-9 联合打包后元素分布

在对任务地点进行打包处理之后，我们需要对会员从经度、纬度、荣誉值等方面进行聚类分析并与任务地点进行关联。

表 5-9 聚集中心

| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 纬度 | 23.34 | 22.94 | 22.12 | 22.36 | 23.65 | 23.29 | 22.56 |
| 经度 | 112.18 | 112.85 | 113.02 | 113.13 | 113.97 | 113.12 | 113.24 |

运用问题二的模型，我们可以得出在每个类的中心点的价格如表 5-10 所示：

表 5-10 每个类的中心点价格

| 类编号 | 1 | 3 | 4 | 5 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|
| 中心点价格 | 71.119 | 69.75 | 75.331 | 78.054 | 73.301 | 76.474 | 79.064 |

分析这种模型对完成情况的影响：

1、将一个范围内的任务联合打包给一个群体的会员，对任务的完成影响为正。

2、定价比原来的定价低，对任务完成度的影响为负。

3、由于在这个模型中考虑到了信誉值的问题，在对会员选择时尽量选择信誉度高的会员，因此对任务完成的影响为正。

4、可能由于打包给一部分会员之后，不能形成良好的竞争机制，对任务完成的效率会有负作用。

## 5.4、问题四模型的建立与求解

### 5.4.1、模型的建立与求解

首先，对新项目位置的数据在地图坐标上找规律，发现项目位置存在位置集中问题。如图 5-10 所示。

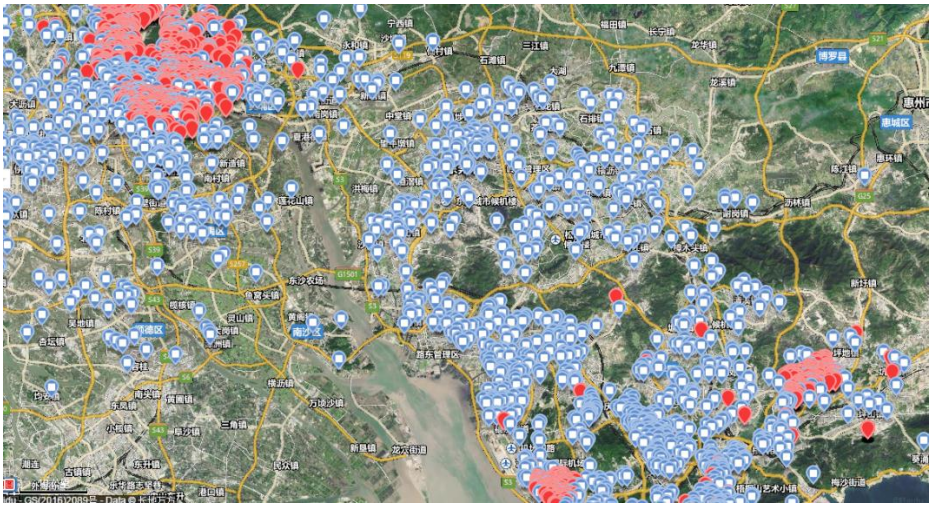

图 5-10 会员和任务位置分布图

图中空心的标记表示会员的信息位置，实心的表示新项目中任务的位置，由图 5-10 可以看出新项目的任务主要聚集在三个地方，为了避免会员对任务的争相选择。

在本问题中我们分别应用上述的三个模型对附件三中的任务给出三种任务定价方案。

1、由问题一中的模型对附件三中数据进行分析。

模型一只考虑任务地点的经纬度和会员的经纬度。运用 SPSS 软件对任务地点和会员位置进行聚类分析

表 5-11 对任务地点和会员等自变量聚类分析

| 编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 纬度 | 22.54 | 23.14 | 23.2 | 22.73 | 22.68 | 22.74 | 23.09 |
| 经度 | 113.93 | 113.52 | 113.16 | 114.24 | 114.16 | 114.38 | 113.33 |
| 编号 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 纬度 | 22.57 | 22.6 | 23.14 | 22.66 | 23.17 | 23.13 | 23.23 |
| 经度 | 113.9 | 114.48 | 113.34 | 114.35 | 113.38 | 113.38 | 113.3 |
| 编号 | 15 | 16 | 17 | 18 | 19 | 20 | |
| 纬度 | 22.71 | 23.2 | 23.16 | 22.59 | 22.79 | 23.18 | |
| 经度 | 113.93 | 113.38 | 113.24 | 113.9 | 114.11 | 113.25 | |

由问题一多元拟合模型：

$$D = 95111 + 8922.3x_1 - 3467.9x_2 - 43.687x_1^2 - 60.943x_1x_2 + 21.39x_2^2 \quad (5\text{-}16)$$

可以计算出各个类中心点任务的定价：

表 5-12 聚集中心的任务定价

| 编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 定价 | 69.166 | 64.547 | 75.236 | 77.924 | 76.276 | 82.762 | 69.65 |
| 编号 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 定价 | 68.519 | 91.851 | 69.152 | 83.992 | 67.692 | 68.097 | 69.616 |
| 编号 | 15 | 16 | 17 | 18 | 19 | 20 | |
| 定价 | 69.545 | 67.297 | 72.278 | 68.679 | 72.626 | 71.852 | |

2、由问题二所得出的模型对附件三中的数据得出任务定价方案，

模型二通过考虑会员的经纬度和任务点的经纬度以及信誉值和会员预定限

额运用 SPSS 进行 K 均值聚类分析[10]。

对任务点的聚类分析如表 5-13，对会员进行经纬度、信誉值、预定限额进行聚类分析：

表 5-13 对任务点的聚类分析

| 编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 纬度 | 22.99 | 113.13 | 22.61 | 21.52 | 33.65 | 22.26 | 29.56 |
| 经度 | 114.02 | 23.03 | 114.47 | 111.17 | 116.97 | 112.8 | 106.24 |
| 编号 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 纬度 | 22.85 | 23.06 | 22.87 | 22.57 | 22.73 | 24.29 | 23.09 |
| 经度 | 112.67 | 113.78 | 113.19 | 114.1 | 114.3 | 116.12 | 112.98 |
| 编号 | 15 | 16 | 17 | 18 | 19 | 20 | |
| 纬度 | 22.82 | 23.12 | 22.8 | 23.28 | 22.56 | 22.98 | |
| 经度 | 113.69 | 113.5 | 114.13 | 113.82 | 113.93 | 113.9 | |

将任务点的经纬度与会员的经纬度进行相关性分析(即他们的马氏距离最短者放在一起)。

由问题二中的多元拟合模型公式：

$$D = 430770 - 5312x_1 - 2680.5x_2 - 9326.8x_3 - 1949.1x_4 + 91.9795x_1^2 + 41.1771x_1x_2 + 99.8465x_1x_3 - 59.1692x_1x_4 - 15.4145x_2^2 + 56.9201x_2x_3 + 33.1496x_2x_4 + 3.1167x_3^2 + 3.9612x_3x_4 - 2.358x_4^2 \quad (5\text{-}17)$$

可以计算出各个类中心的任务定价

表 5-14 各类聚集中心的任务定价

| 编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| 定价 | 79.138 | 92.13 | 67.047 | 79.125 | 79.68 | 64.758 | 71.47 |
| 编号 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 定价 | 71.146 | 71.973 | 74.496 | 70.184 | 68.586 | 67.955 | 68.039 |
| 编号 | 15 | 16 | 17 | 18 | 19 | 20 | |
| 定价 | 68.897 | 66.544 | 69.307 | 78.965 | 74.806 | 67.535 | |

3、由问题三得出的模型处理附件三中得出一个任务定价方案

模型三通过对一些比较集中的数据进行打包处理，筛选出那些不被打包的数据。我们运用 SPSS 先对任务地点进行 K 均值聚类分析：

表 5-15 对任务地点进行 K 均值聚类分析观察值数目

| 每一個聚集中的观察值數目 | | |
|---|---|---|
| 聚集 | 1 | 103.000 |
| | 2 | 2.000 |
| | 3 | 3.000 |
| | 4 | 6.000 |
| | 5 | 1.000 |
| | 6 | 3.000 |
| | 7 | 241.000 |
| | 8 | 2.000 |
| | 9 | 9.000 |

| | | |
|---|---|---|
| | 10 | 2.000 |
| | 11 | 21.000 |
| | 12 | 410.000 |
| | 13 | 244.000 |
| | 14 | 443.000 |
| | 15 | 1.000 |
| | 16 | 163.000 |
| | 17 | 2.000 |
| | 18 | 2.000 |
| | 19 | 224.000 |
| | 20 | 184.000 |
| 有效 | | 2066.000 |
| 遗漏 | | .000 |

由上表可以看出在分 20 类时有几类的容量很小，因此这些数据不宜成为打包数据，即可以分为 8 类：

表 5-16 打包结果

| 编号 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 纬度 | 22.71 | 22.58 | 23.09 | 23.14 |
| 经度 | 114.24 | 113.89 | 113.31 | 13.38 |
| 编号 | 5 | 6 | 7 | 8 |
| 纬度 | 23.14 | 22.64 | 22.64 | 23.17 |
| 经度 | 114.28 | 113.84 | 113.84 | 113.25 |

在对会员的数据处理时，重点考虑信誉值后对其进行 K 均值聚类分析：

表 5-17 考虑信誉聚类结果

| 编号 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 纬度 | 24.8 | 113.13 | 24.29 | 22.48 |
| 经度 | 113.61 | 23.03 | 116.12 | 113.91 |
| 编号 | 5 | 6 | 7 | 8 |
| 纬度 | 33.65 | 20.34 | 29.56 | 27.12 |
| 经度 | 116.97 | 110.18 | 106.24 | 111.02 |

在通过会员与任务进行关联后，以模型二的公式计算出各类中心任务点的定价。

表 5-18 模型二求出的聚类中心点定价

| 编号 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 定价 | 67.252 | 71.051 | 75.942 | 68.888 |
| 编号 | 5 | 6 | 7 | 8 |
| 定价 | 70.746 | 70.463 | 70.19 | 71.201 |

对三个模型进行对比分析：

表 5-19 模型对比

| | 模型一 | 模型二 | 模型三 |
|---|---|---|---|
| 相关系数 | 0.86625 | 0.992 | 0.9935 |
| F 检验 | 5.1815 | 8.24 | 17.553 |
| 考虑方面 | 任务点的经纬度 | 任务点与会员的经纬度、信誉值、预定限额 | 任务点与会员的经纬度、信誉值、预定限额 |
| 分析方法 | 聚类分析、多元拟合 | 聚类分析、多元多次拟合 | 聚类分析、多元多次拟合、数据的筛选和打包 |
| 任务完成率 | 低 | 中 | 高 |

由上表可知对于附件三中的任务定价方案模型三的相关系数与 F 检验都比模型一和模型二好，因此在对附件三任务定价时尽量运用模型三去解决。

# 六、模型评价和推广

## 6.1 模型的评价

### 6.1.1 模型的优点

1）本文通过使用 MATLAB、Excel 和 SPSS 等处理数据的工具，在大数据中提炼出有用的数据，使模型的建立更加的方便。

2）通过对任务定价的因素进行多方面的分析，建立了定价与任务位置和会员位置等因素的数学模型，使得模型简单可行。

### 6.1.2 模型的缺点

1）由于数据的不足，模型的建立没有考虑到会员完成任务的成本、发任务方需要利益最大化等问题，任务的定价问题会受到其的影响。

2）题中的数据只是以广东省的数据进行建立的，没有考虑到不同城市的文化、经济发展形式和居民的生活水平差异，模型具有一定的片面性。

## 6.2 模型的推广

本文是通过任务的位置与会员信息建立的关系，建立对任务定价的方案。这是从接任务者与任务的关系来建立的模型，也可以从经济学角度考虑定价方案，即任务发布者[7]与任务执行者之间的质量、利润关系，如图 6-1 所示。

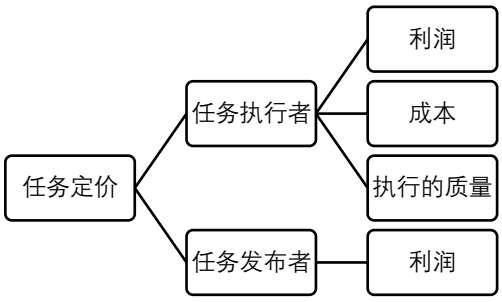


图 6-1 定价与任务发布者与任务执行者之间关系

任务定价在众包平台有着举足轻重的位置，合理的任务定价能有效的促进任务的完成。任务定价与众多因素有关，建立数学模型需要更为全面的考虑众多因素。任务定价的模型可以推广到多个领域，进行应用。比如：网络竞赛、企业的设计招标和政府的众包；只要定价合理，大众的力量就能得到更好的利用。

# 七、参考文献


[1]孙信昕.众包环境下的任务分配技术研究[D].扬州大学,2016.

[2]陈家银.猪八戒众包平台数据分析与众包模式设计[D].大连理工大学,2016.

[3]郝琳娜,侯文华,张李浩,刘猛.基于众包虚拟社区的诚信保障和信誉评价机制研究[J].系统工程理论与实践,2014,3411:2837-2848.[2017-09-17].

[4]孟韬,张媛,董大海.基于威客模式的众包参与行为影响因素研究[J].中国软科学,2014,12:112-123.[2017-09-17].

[5]陈晓,赵晶玲.大数据处理中混合型聚类算法的研究与实现[J].信息网络安全,2015,04:45-49.[2017-09-17].

[6]冯剑红,李国良,冯建华.众包技术研究综述[J].计算机学报,2015,3809:1713-1726.(2014-11-24)[2017-09-17].http://kns.cnki.net/kcms/detail/11.1826.tp.20141124.1316.001.html

[7]刘晓钢.众包中任务发布者出价行为的影响因素研究[D].重庆大学,2012.

[8]向泽君,徐占华,饶鸣,龙川.利用数据打包发布海量栅格瓦块地图的方法[J].测绘通报,2014,06:75-78.[2017-09-17].DOI：10.13474/j.cnki.11-2246.2014.0196

[9]白雪.聚类分析中的相似性度量及其应用研究[D].北京交通大学,2012.

[10]许丽利.聚类分析的算法及应用[D].吉林大学,2010.


# 八、附录

## 拟合源代码

```
y=xlsread('fujian1','t_tasklaunch','D2:D836');%  读取附件一中任务价格
y=y';
X1=xlsread('fujian1','t_tasklaunch','B2:B836');% 读取附件一中任务位置纬度
X1=X1';
X1=xlsread('fujian1','t_tasklaunch','C2:C836');% 读取附件一中任务位置经度
X2=X1';
format short g
Y=y'
X11=[ones(1,length(y));X1;X2]'
B1=regress(Y,X11)% 多元一次线性回归
[m,n]=size(X11)
X22=[];
for i=2:n
    for j=2:n
        if i<=j
            X22=([X22,X11(:,i).*X11(:,j)]);
```

```
        else
        continue
        end
    end
end
X=[X11,X22];
[B2,bint,r,rint,stats]=regress(Y,X)% 多元二次回归
[Y X*B2 Y-X*B2]
plot(Y,X11*B1,'o',Y,X*B2,'*')
hold on,
line([min(y),max(y)],[min(y),max(y)])
axis([min(y) max(y) min(y) max(y)])
legend('一次线性回归','二次线性回归')
xlabel('实际值');
ylabel('计算值')
```

由于其它的拟合代码类似，只需要自行调整部分即可；

绘图代码

任务所在地的图形分布代码：

```
X1=xlsread('fujian1','t_tasklaunch','B2:B836');% 读取附件一中任务位置纬度
X1=X1';
X1=xlsread('fujian1','t_tasklaunch','C2:C836');% 读取附件一中任务位置经度
X2=X1';
Plot(X1,X2,' R*' )
```

会员所在地的分布情况

```
X1=xlsread('fujian2','t_tasklaunch','B2:B836');% 读取附件一中任务位置纬度
X1=X1';
X1=xlsread('fujian2','t_tasklaunch','C2:C836');% 读取附件一中任务位置经度
X2=X1';
Plot(X1,X2,' R*' )
```

会员和任务同时分布图的代码

```
X1=xlsread('fujian1','t_tasklaunch','B2:B836');% 读取附件一中任务位置纬度
X1=X1';
X1=xlsread('fujian1','t_tasklaunch','C2:C836');% 读取附件一中任务位置经度
X2=X1';
Plot(X1,X2,' R*' )
Hold on
t=xlsread('fujian1','t_tasklaunch','B2:B836');% 读取附件一中会员位置纬度
t=t';
z=xlsread('fujian1','t_tasklaunch','C2:C836');% 读取附件二中会员位置经度
z=z';
Plot(z,z,' m*' )
```

其它绘图和以上代码类似。