讨论众包任务的成功因素与优化模型

摘要

本题主要研究众包经济模式下对于任务定价、分配的优化策略。使用曲面拟合的方法，建立核回归模型和回归模型，并提出新的定价模型；我们建立了完成度检验模型，通过这个模型，可以得出一定区域内任务的完成数量，也可以衡量一件任务完成的可能性，以此来评价定价方案的合理程度；对任务打包，主要得出当任务完成度最高时的区域大小。

对于问题一，首先，采用函数拟合的方法求解出定价和其他因素的关系，在众多拟合方式中，常用的最小二乘法因需要提前确定函数形式而并不适用，在这里采用核回归的方式更为合理。之后，将已完成任务和未完成任务分开，依照上述算法分别计算，用 MATLAB 拟合出二者的函数图像及解析式，在这些图像解析式建立并求解后可以得出：未完成的原因有平均定价过低、会员人数过少、任务数量和会员数量的比例过高等。

对于问题二，首先，我们在核回归模型的基础上建立回归预测模型，仅对附件一中已完成的任务进行拟合，求得函数图像，从而得出一系列可能的定价数值，记为定价 1，再利用附件一未完成任务的数据对定价 1 进行修正，得到定价 2，使用完成度检验模型对其进行检验，以此确定这个定价方案的合理程度。然后，建立完成度检验模型，模型求解原理是通过求出全部区域及其分块（在这里实际操作将 80%的区域分为了 10*10 区域小块）的相关量，即一定区域内任务总价值/任务数量和任务总数量/会员总数量，再用这些数据进行拟合求出函数关系式。最后将剩余 20%区域分块，带入此函数关系式来检验其合理性。使用此模型可分析问题一中未完成原因，并可以作为重要依据来衡量之后几个问题中求出的定价方案合理性。

对于问题三，可对附件一中已有数据进行多次平均分块和非平均分块，每次分块大小并不一致但依照一定规律，在多次计算后得出每次分块后的平均任务完成度，以此寻找出一个任务包最合适的面积。同样的方式，确定具体一定区域内任务总数量和任务数/会员数的合理值。这相当于多次尝试、优化后寻找确定区域内的较优解，一步步修改之前的定价模型。

对于问题四，可综合前三问中所有因素进行综合分析和考察，利用已有模型求解出定价方案并根据以往的未完成数据进行修正，在限制条件下给出更为合适的结果。最终我们拿出的定价方案经模拟检验，在不进行打包的情况下，总完成度约为 0.896425，这个完成度明显大于原有方案的 0.625150，且经过打包优化后，这个完成度将会更高。

关键词：众包 优化 核回归 回归预测 完成度检验 曲面函数拟合

一、问题重述

1.1 背景

众包又称为网络化社会生产，是指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，具有低成本生产、联动潜在生产资源、提高生产效率、而且满足用户个性化需求等优势。移动互联网和人工智能的迅猛发展，颠覆了很多传统行业。随着第三产业的突飞猛进和用工成本的逐年提高，新一代的 90 后更倾向于自由度更高的工作，兼职行业逐渐成为一个潮流。特别是对于房产、教育、快消品等具有周期性用工需求的企业，用兼职员工替代全职员工是最经济的一种做法。汤劲武在 2015 年创建了拍拍赚，自主研发 AI 技术，并采用众包模式，实现实体零售渠道监测的智能化，最终达成实体零售大数据化的目标。在众包模式下，任务发布者和众包平台需要找到以更小的代价（在保证任务完成度的情况下使得总的定价尽量小）完成任务的方式。这就需要制定更合理的定价策略和分配模式，在定价上，影响因素主要是任务和会员所在地的位置及其重合程度，一定区域内总的任务数量、总的会员数量及其比例，及任务的难易程度等。而在此基础上，打包模式不仅可以缓解由于一定区域内任务接受者争相选择的情况，还能够在打包的时候对定价进行调整，一方面对于任务密集区、易于完成的任务的定价可以适当调低，降低了任务发布者的总金额支出的同时也使得任务接受者在更小的区域完成更多的任务从而降低任务完成成本；另一方面对于任务稀疏区域、任务难度大的地方，可适当调高价格，这增加了任务完成度也使得任务接受者可获得更高的报酬。经过二次优化之后，众包模式更加具备合理性和可操作性。

1.2 问题产生与研究意义

第一、在传统模式下，某个企业如果要对全国性的产品进行铺货率调查，会委托给市场调查公司。调查公司再把一个项目分包给几十个区域性的数据采集公司，后者开始设计调查问卷后，接着再雇大量兼职人员在大街上拦截访问，或者电话访问，一份 2000 人的问卷样本通常需要三个月才能完成。这种传统模式效率低、质量差，且无实时改进预警机制的缺点。

第二、企业难以找到需即用、低成本地合适的短期人员。招聘网上信息量巨大，通过中介找兼职还要经过层层代理，成本较高。

第三、兼职行业中的主体乱、规则乱、没品牌，甚至存在大量虚假、骗人信息，对兼职人员来说，难以即时地找到附近且靠谱的兼职。且对大学生而言，难以在兼顾学业的同时找到合适的兼职。

研究意义方面，众包已经成为一种新型的电子商务模式并占据着越来越重要的市场地位。众包是一种开放式创新，其成功与否与任务发布者的出价密切相关，研究任务的出价策略对于任务执行者完成任务积极性和任务发布者的投资具有重要意义。

1.3 问题的提出

附件一是一个已结束项目的任务数据，包含了每个任务的位置、定价和完成情况（“1”表示完成，“0”表示未完成）；附件二是会员信息数据，包含了会员的位置、信誉值、参考其信誉给出的任务开始预订时间和预订限额，原则上会员信誉越高，越优先开始挑选任务，其配额也就越大（任务分配时实际上是根据预订限额所占比例进行配发）；附件三是一个新的检查项目任务数据，只有任务的位置信息。请完成下面的问题：

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。

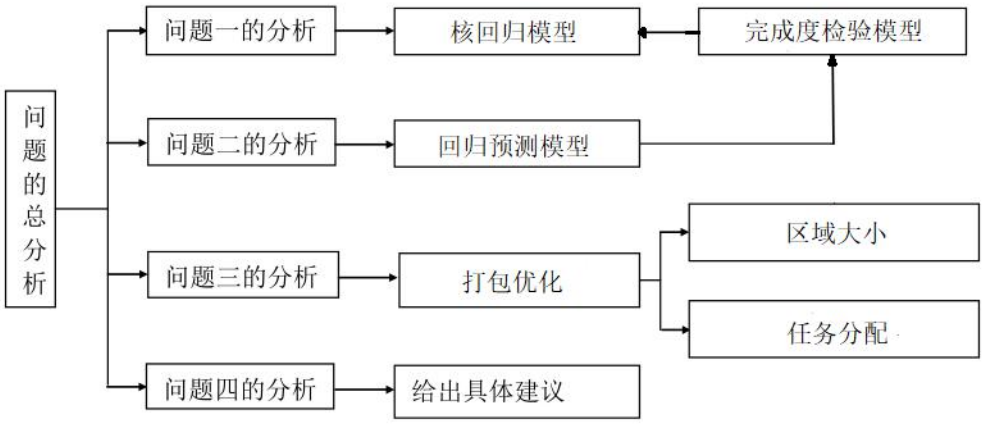
3. 实际情况下，多个任务可能因为位置比较集中，导致用户会争相选择，一种考虑是将这些任务联合在一起打包发布。在这种考虑下，如何修改前面的定价模型，对最终的任务完成情况又有什么影响？

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。

二、问题分析

本文研究劳务众包的定价方案。由于将工作做外包给社会上陌生人员，无法保证任务一定会被执行，从接受任务者的角度来说，是否会完成此项任务的唯一标准就是自身完成任务所投入的精力与完成任务所得金额的大小关系，从任务发布者的角度来说，任务的完成程度仅仅是与该任务的标价有关，标价越高，那么任务被完成的可能性就越大。另一方面如果任务标价过高会是任务发布者收益受限。总的来说，如果任务发布者提前通过任务的某些参量得到该任务应该分配的合适金额数就会使得任务具有较高的完成度的同时保证任务发布者的收益。定价需在双方都可接受的的一个范围里才可使交易完成。定价由很多因素决定，如会员与目的地之间的距离，在某一区域内会员的密集程度等。

对这个问题的分析流程图如下：



图表 1

2.1 问题一的分析

从该视角出发的研究将发布者和威客的决策看作是不完全信息动态博弈问题。研究的目标是最大化任务发布者的收益。附件一中给出会员所在位置的经纬度、任务标价和任务执行情况。为研究出价规律，建立核函数研究标价与经纬度的关系。

2.2 问题二的分析

继承问题一中的模型，但仅通过对附件一中已完成的任务进行拟合，求得函数解析式和图像，然后使用该表达式得出一系列可能的定价数值。最后通过建立完成度检验模型比较这个定价方案与问题一中原有定价方案的合理程度。

2.3 问题三的分析

从该视角出发的研究主要分析任务之间的竞争以及任务解决者之间的竞争对出价的影响。打包之后的区域面积、任务和会员数量及比例、平均定价、会员信誉等级等因素对最终的任务完成情况都可能产生影响，我们利用优化模型一步步一个个因素进行考察分析，以此修改之前的定价模型。

三、模型的假设

为了使得问题得到简化并使其容易求解，在考虑实际情况与所求解问题所要求的精度情况下，特此做出如下合理假设：

1、不考虑任务的难易程度，即不考虑任务本身内容对完成情况的影响；
2、不考虑突发、不可控和其他未知因素导致的结果偏差甚至错误。
3、数据量规模足够大，具有很好的统计回归意义；

四、模型建立与求解

4.1 问题一模型建立与求解

4.1.1 基于核回归模型对定价规律分析

（1）模型的准备

对于附件一所给数据，建立定价金额与经纬度的关系函数。在本题中我们需要对散乱点进行三维连续曲面拟合，得到曲面 $z = f(x,y)$，其中 z 代表原有定价，x 和 y 分别代表经纬度。因不了解曲面方程的具体形式，所以这里对于指定形式求系数的最小二乘法并不适用。我们采用了核回归方法，这是一种非参数回归方法，即不对 $z = f(x,y)$的形式做任何假定，属于散点曲面重构方法之一。其采用局部加权方法，用点 x 附近的Y_i的加权平均表示 $z = f(x,y)$，权重为核函数的值，而邻域由核函数的宽度控制。[1]具体建模过程如下：

记 $v = [x,y]^T$，记曲面表达式为

$$z = f(x,y) = f(v). \tag{1}$$

对其进行二阶 Taylor 展开：

$$f(v + t) = f(v) + t^T \acute{y}_f(v) + \frac{1}{2} t^T H_f(v) t + o(t^T t). \tag{2}$$

其中用$\acute{y}$和 H 标记梯度算子:

$$y_f(v) = [\frac{\partial f(v)}{\partial x} \ \frac{\partial f(v)}{\partial y}]^T. \tag{3}$$

$$H_f(v) = \frac{1}{2}[\frac{\partial^2 f(v)}{\partial x^2} \ 2\frac{\partial^2 f(v)}{\partial x \partial y} \ \frac{\partial^2 f(v)}{\partial y^2}]^T. \tag{4}$$

记 vech 为上三角矩阵的向量化：

$$\mathrm{vech}\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = [a \ b \ d]^T. \tag{5}$$

这里，采用核回归方法使得回归系数向量$\{\beta_n\}_{n=0}^{p}$让 Bias（偏差）式(6)最小：

$$Bias\{\hat{f}(v)\} = \min_{\{\beta_n\}_n^p} \sum_{i=1}^{n} \{Y_i - \beta_0 - \beta_1^T(V_i - v) - \beta_2 vech\{(V_i - v)(v_i - v)^T\} - \cdots\}^2 K_H(V_i - v). \tag{6}$$

记 n 个离散点为：

$$Y=[y_1\ y_2 \ldots y_n]^T. \tag{7}$$

记$\{\beta_n\}_{n=0}^p$的向量化为：

$$B=[\beta_1\ \beta_1^T \ldots \beta_n^T]^T. \tag{8}$$

记权重矩阵为：

$$\boldsymbol{W} = \begin{pmatrix} K_H(\boldsymbol{V}_1 - \boldsymbol{v}) & 0 & \cdots & 0 \\ 0 & K_H(\boldsymbol{V}_2 - \boldsymbol{v}) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & K_H(\boldsymbol{V}_n - \boldsymbol{v}) \end{pmatrix}. \tag{9}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & (\boldsymbol{V}_1 - \boldsymbol{v})^{\mathrm{T}} & vech^{\mathrm{T}}\{(\boldsymbol{V}_1 - \boldsymbol{v})(\boldsymbol{V}_1 - \boldsymbol{v})^{\mathrm{T}}\} \\ 1 & (\boldsymbol{V}_2 - \boldsymbol{v})^{\mathrm{T}} & vech^{\mathrm{T}}\{(\boldsymbol{V}_2 - \boldsymbol{v})(\boldsymbol{V}_2 - \boldsymbol{v})^{\mathrm{T}}\} \\ \vdots & \vdots & \vdots \\ 1 & (\boldsymbol{V}_n - \boldsymbol{v})^{\mathrm{T}} & vech^{\mathrm{T}}\{(\boldsymbol{V}_n - \boldsymbol{v})(\boldsymbol{V}_n - \boldsymbol{v})^{\mathrm{T}}\} \end{pmatrix}. \tag{10}$$

根据上式，则有：

$$\hat{B} = (X^T W X)^{-1} X^T W Y. \tag{11}$$

由 Nadaraya-Watson 估计可得出，高斯核回归方程为：

$$\hat{f}(v) = \frac{\sum_{i=1}^{n} K_H(V_i - v) y_i}{\sum_{i=1}^{n} K_b(V_i - v)}. \tag{12}$$

（2）模型的求解

本题中，我们使用了 MATLAB 进行计算。使用核回归方法编写代码(见附录程序一)，对于附件一中给定的定价 z，经纬度 x、y，录入离散三维数据点数据并进行拟合，为获得更好的 SSE 和 R-square 的值，我们采用多项式拟合的方法，求出函数解析式极其参数：

Linear model Poly42:

f(x,y) = p00 + p10*x + p01*y + p20*x^2 + p11*x*y + p02*y^2 + p30*x^3 + p21*x^2*y + p12*x*y^2 + p40*x^4 + p31*x^3*y + p22*x^2*y^2

where x is normalized by mean 22.98 and std 0.2453

and where y is normalized by mean 113.5 and std 0.3729

Coefficients (with 95% confidence bounds):

p00 =	66.64	(66.28, 66.99)
p10 =	-1.809	(-2.352, -1.266)
p01 =	0.8197	(0.5126, 1.127)
p20 =	1.508	(1.096, 1.92)
p11 =	3.854	(3.281, 4.427)
p02 =	2.248	(1.935, 2.56)
p30 =	0.3705	(0.07391, 0.6671)
p21 =	-0.9419	(-1.468, -0.4164)
p12 =	1.131	(0.7193, 1.543)
p40 =	-0.1593	(-0.2442, -0.07438)
p31 =	-0.2064	(-0.4279, 0.01504)
p22 =	0.8977	(0.6576, 1.138)

Goodness of fit:

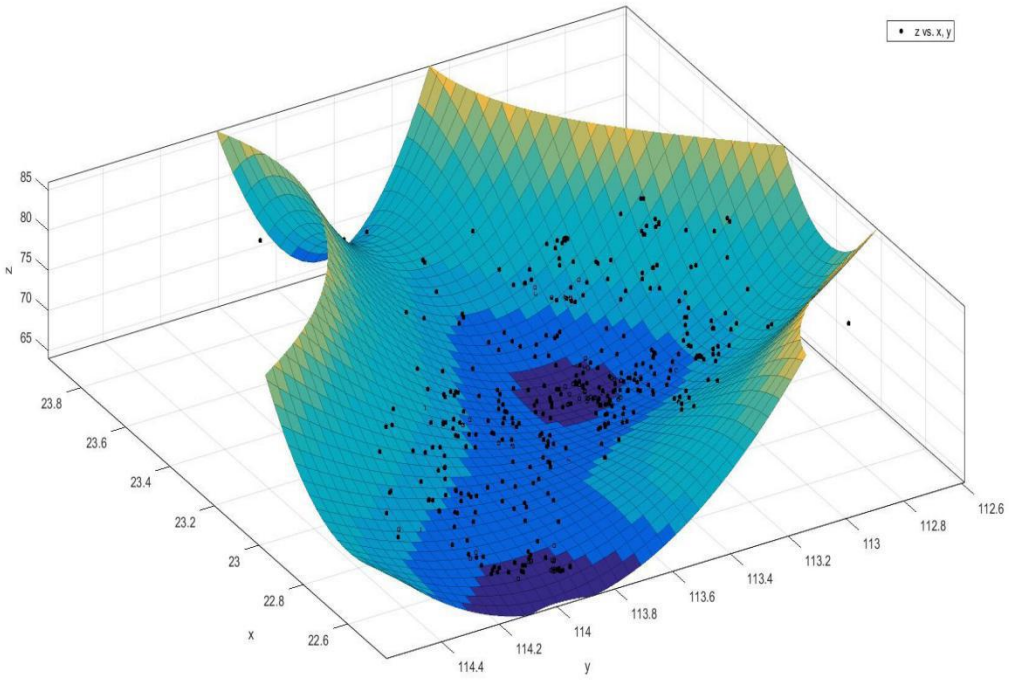
SSE: 19.12

R-square: 0.6651

Adjusted R-square: 0.6606

RMSE: 2.629

得到的连续曲面函数图像如图表 2：



图表 2 已结束任务的价格-GPS 拟合曲面

图中颜色越深表示价格越低，黑点表示已结束任务数据点，分析点的分布我们可以看出大部分任务位置集中在图形中最深色和稍次之的区域内，这些任务对应的定价也大致处于最低点状态，而对于在两个低谷外围的任务点来说，其价格呈现高速上涨的趋势。即，在任务密集的地方价格较低，而较偏离任务密集区的地方，价格会升高。

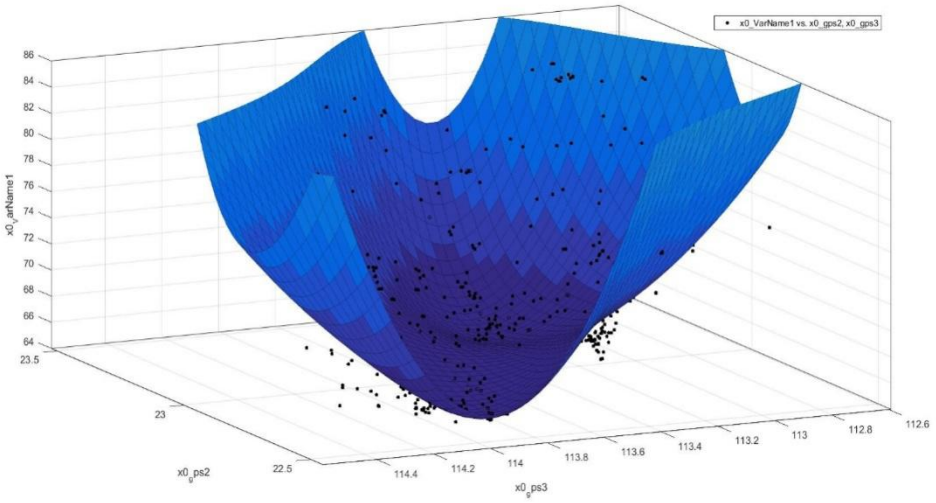
结合函数分析可知，定价以（23.2，113.4）和（22.6，114）为谷底，呈现一种双谷底形状，越往四周扩散价格越高，且梯度逐渐增大。根据会员分部图和任务分布图叠加可知，未完成任务大多分布于会员数量少的位置

4.1.2 分析未完成任务原因

将附件一中未完成数据点用核回归的方法在 Matlab 中进行曲面拟合，如图表 2，我们可以发现未完成的任务数主要集中在价格越低的部分（颜色较深的部分）和位置较偏的地方。由此，我们可以得出：

1、在相对偏僻的地方，任务完成度小，主要是会员人数少，且实际情况可能会有交通不便等因素的影响，对任务发布者来说，可能不愿出很高的价格，对会员来说，花费成本（时间成本+实际成本）不能带来更大收益。

2、在图中也可看出，价格低的区域完成度低。在这些区域内是任务机会多，若价格较低，则难以发动接受者的完成任务的热情。



图表 3 未完成任务的价格-GPS 拟合曲面

图表 3 对比图表 2 可知，要使任务完成，价格需要在任务发布者和执行者都可接受的一个范围内；且任务地点若较为偏远，也会影响其完成度。

4.2 问题二模型建立与求解

4.2.1 基于回归预测模型设计定价

（1）问题分析

为保证附件一中已完成的任务在新模型下仍是可完成的状态，只需未完成任务上重新设计定价的思路。于是在本问中，只保留第一问中顺利完成的任务数据，利用回归分析模型，求解表达式，见公式（17），并绘制函数图像，如图表 2。之后将未完成的数据点代入公式（17）进行预测，并在得到预测值后与未完成任务数据进行对比，进行后期修正。再使用未完成任务的数据与初步得到的定价数据进行对比，将初步得到的价格比未完成任务价格还低的数值提升 x%，其中 x 的确定方式是根据完成度检验模型得到的函数。经过一系列修正之后，得到精度更高、符合度更好的定价方案。

（2）模型准备

该回归模型原理是：

对 z 进行一个非线性变换 ψ，f(z)是变换结果的线性函数：

$$f(z)=w^T \psi(z) \tag{13}$$

w 由训练样本的非线性变换 ψ(xi)的线性组合构成：

$$w=\sum i \alpha_i \psi(x_i) \tag{14}$$

两者结合，得到完整形式：

$$f(z)=[\sum i\alpha_i\psi(x_i)]\,\psi(z)=\sum i[\alpha i\psi(x_i)\,\psi(z)] \quad (15)$$

记 $\kappa(xi,xj)=\psi(xi)\,\psi(xj)$，称为核函数。则：

$$f(z)=\sum i\alpha_i\,\kappa(z,xi)=\alpha T\kappa(z) \quad (16)$$

α 为 N×1 矢量，$\kappa(z)$ 为 N×1，第 i 个元素为训练样本 xi 和测试样本 z 的核函数值。f(z)是关于 z 的非线性函数，但却是关于 $\kappa(z)$ 的线性函数，可以使用线性函数的优化方法求解 α。原始参数 w 处于原空间 prime space 中，新参数 α 则处于对偶空间 dual space 中。[5]

（3）模型求解

本次选取点为附件一中所有完成任务的地理信息和定价，利用 MATLAB 求解回归函数拟合曲面（见）并绘制图像，结果如下：

Linear model Poly42:

f(x,y) = p00 + p10*x + p01*y + p20*x^2 + p11*x*y + p02*y^2 + p30*x^3 + p21*x^2*y + p12*x*y^2 + p40*x^4 + p31*x^3*y + p22*x^2*y^2

where x is normalized by mean 23.02 and std 0.2226

and where y is normalized by mean 113.5 and std 0.3491

Coefficients (with 95% confidence bounds):

p00 =	66.06	(65.75, 66.38)
p10 =	-2.308	(-2.825, -1.792)
p01 =	1.052	(0.7893, 1.314)
p20 =	1.617	(1.3, 1.933)
p11 =	1.757	(1.261, 2.252)
p02 =	2.516	(2.263, 2.769)
p30 =	0.325	(0.1041, 0.5459)
p21 =	-0.7733	(-1.121, -0.4259)
p12 =	1.484	(1.138, 1.83)
p40 =	-0.1316	(-0.1853, -0.07791)
p31 =	-0.1323	(-0.3107, 0.04603)
p22 =	0.07609	(-0.1531, 0.3052)

Goodness of fit:

SSE:10.28

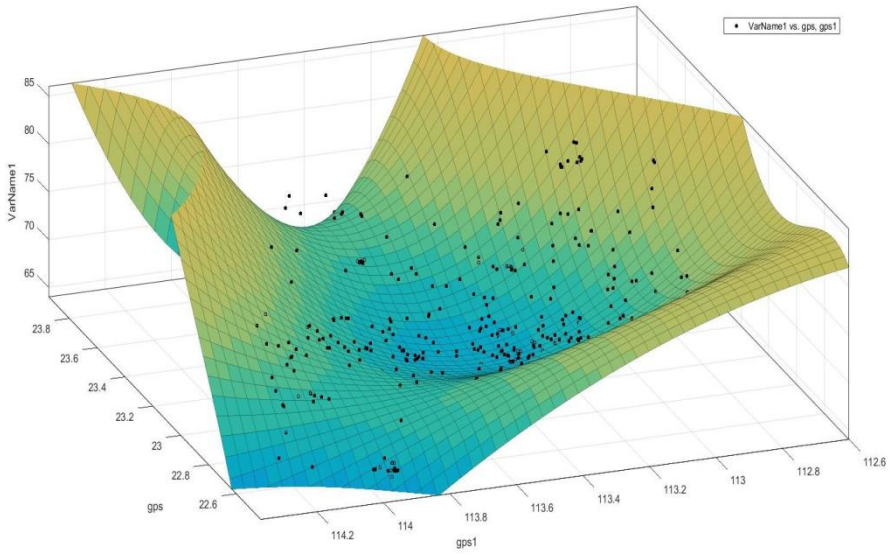
R-square: 0.8323

Adjusted R-square: 0.8287

RMSE: 1.994

得出曲面函数模型是

f(x,y)=p00+p10*x+p01*y+p20*x^2+p11*x*y+p02*y^2+p30*x^3+p21*x^2*y+p12*x*y^2+p40*x^4+p31*x^3*y+p22*x^2*y^2. (17)



图表 4 已完成任务价格-GPS 拟合曲面

为检验此模型与原模型的优劣程度，根据 Terwiesch 等人的如下模型：

$$v = \rho \max_{i=1\ldots n}\{v_i\} + （1-\rho）\frac{\sum_{i=1\ldots n} v_i}{n}$$

我们提出完成度检验模型，具体如下

4.2.2 完成度检验模型

Laursen 和 Salter（2005）的研究结果表明当企业有足够的能力来开发他们的外部资源时，拥有更多的外部资源将会增加他们 R&D 的创新绩效。Chesbrough（2003）同样认为一个开放式创新过程中的核心是寻找具有商业潜力的外部资源。在比赛当中，如果出现了更多的参与者，那么更多的外部资源将会被开发出来，因此任务发布者也能够得到更好的价格和完成质量。基于以上分析我们提出假设，即会员数密度和标价对任务完成情况有正方向的影响。

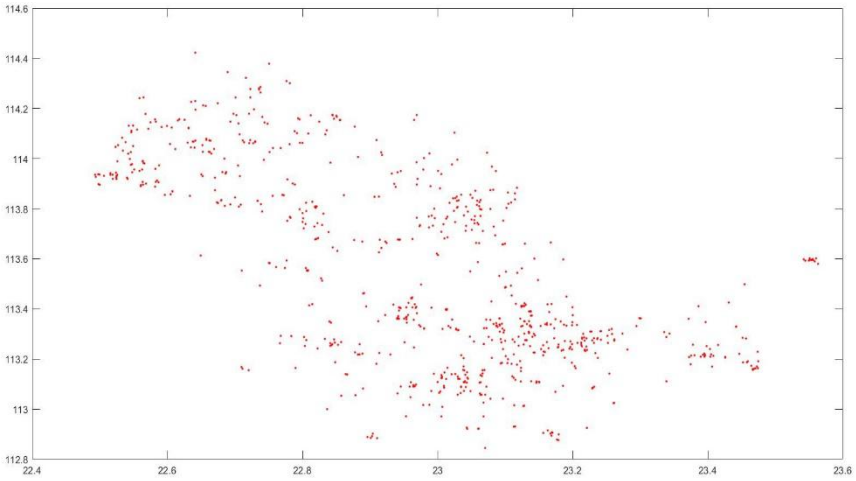
为了衡量一个定价方案内具体一件任务的完成的可能性和定价金额和任务难易程度的关系，将所在范围分为 10*10 小区域，求得每个区域内任务总数量和会员总数量及其比例（稀缺性）、已完成任务数与总任务数量的比例（完成度），可以知道，完成度和稀缺性、平均价格有着密切关系。设小区域内任务已完成数量（m）和该区域内总任务数量（M）的比例（即完成度）为 P，一个小区域内的所有任务的平均定价为 R,一个小区域任务总数量为 N，一个小区域会员总数量为 n。则 P=F(m/M)=f(R,n/N)，求解过程如下：

对附件一中所有区域进行划分，得到 10*10 区域，进行区域筛选，去除没有任务或者任务极少的区域。求得每个可用的小区域的会员数量、任务总数、完成任务数量、完成度、任务总定价、任务均价，使用 MATLAB 求解此二元函数解析式及图像。得到函数为

f(x,y) = p00 + p10*x + p01*y
Coefficients (with 95% confidence bounds):
p00 = -0.006036 (-0.03932, 0.02725)
p10 = 0.01427 (0.01358, 0.01496)
p01 = 0.01746 (-0.004279, 0.03921)：

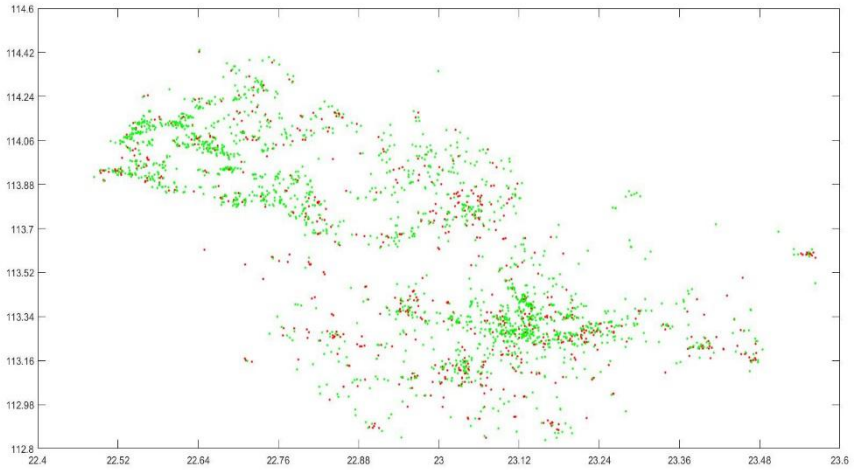
对于附件二中的会员地理位置信息，也应给予一定权重的考虑。具体方式如下：

选取点为附件一中所有完成任务的地理信息和定价，利用 MATLAB 求解回归函数拟合曲面并绘制图像。



图表 5 已完成任务价格-GPS 平面图

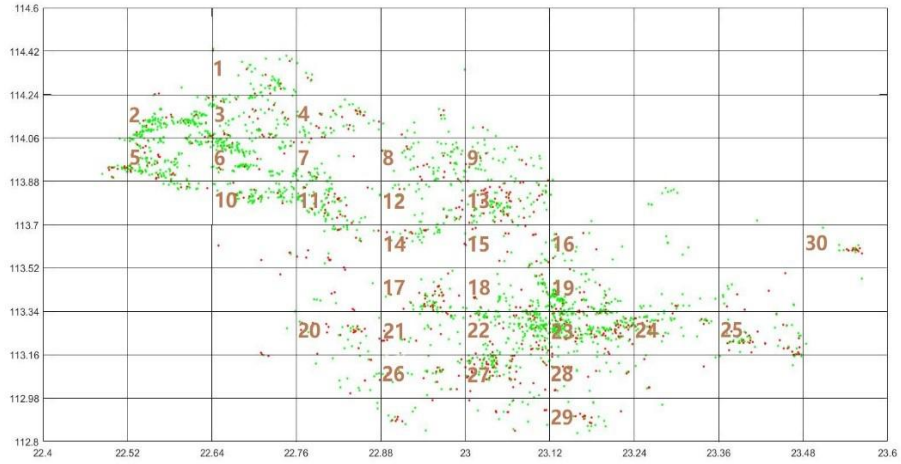
单单看此图无法得出正常结论，因此我们将附件二中每位会员所在位置的 GPS 点放入此图当中。为了使数据更加集中，图形更加明晰，我们删去了少部分坐标位置极其偏离总体集中位置的点，最终得到了下面的图，其中绿色代表会员所在地，红色代表任务所在地，同时为了方便后续计算，我们将 x 轴和 y 轴分出 11 个坐标点



图表 6 添加会员 GPS 平面图

从图中我们可以看出，任务分布的大致走向与会员所在地较为吻合。以纬度 22.52 到 22.76 经度 113.7 到 114.24 和纬度 23.12 到 23.24 经度 113.16 到 113.34 的两个区域较为集中，任务的分布不会偏于会员分布之外。

为了定量研究会员分布与任务分布的关系，我们将上图中的全部区域分成 10*10 的 100 个方块，再对可以参与计算的方块进行筛选，去掉没有任务点的方块、任务点和会员分布极为有限的方块，最终选取了 30 个方块参与计算，计算顺序详见下面方块的编号排序，如图表 7.



图表 7 分区图

对于方块的计算我们首先选出方块中全部任务数量记作 tasklength，被完成的任务的个数几座 finishlength，二者比值算出该方块实际的完成度记作因变量 z，即

$$z = \frac{tasklength}{finishlength}$$

接着累计该方块所有任务的价值，得到总价值 value 再用其除以该方块的任

务的个数得到该方块的任务的平均价值记作自变量 x，最后用总任务数除以该方块的总人数 peoplelength 得到第二个自变量 y，即

$$x = \frac{value}{tasklength}$$

$$y = \frac{tasklength}{peoplelength}$$

详见下表:

表格 1 完成度

方块编号	总价/任务个数 --X	任务数/会员数--Y	实际完成度--Z
1	43.6	0.2703	0.6
2	32.8333	0.1622	0.4583
3	40.2593	0.3913	0.5556
4	45.35	0.5405	0.65
5	48.3415	0.3504	0.6585
6	42.4474	0.152	0.5789
7	39.7143	0.2692	0.5714
8	31	0.2778	0.4667
9	37.8333	0.2727	0.5556
10	48.5	0.2647	0.6667
11	41.0962	0.3768	0.6154
12	24.3636	0.3438	0.3636
13	65.5577	0.7324	0.9423
14	27.375	0.5455	0.4167
15	65.2143	0.9333	0.9286
16	43.6	0.3571	0.6
17	44.1579	1.0857	0.6579
18	63.675	0.3279	0.95
19	39.4259	0.2571	0.5556
20	39.65	1.0526	0.6
21	36.7308	0.5417	0.5385
22	53.0638	0.3264	0.766
23	30.3667	0.4972	0.4333
24	27.8421	0.4524	0.3684
25	39.2286	1.0606	0.5429
26	40.9375	0.7619	0.625
27	52.2632	1	0.7544
28	51.2	1	0.7333
29	41.3	0.7692	0.6
30	14	1.75	0.2143

通过基础的多项式拟合我们得到了吻合度较高的结果：

Linear model Poly11:

f(x,y)=-0.006036+ 0.0182068x- 1.049737y^2

Coefficients (with 95% confidence bounds):

p00 =	-0.006036	(-0.03932, -0.02725)
p10 =	0.0182068	(0.01758, 0.01996)
p01 =	- 1.049737	(-1.054279, -1.03921)

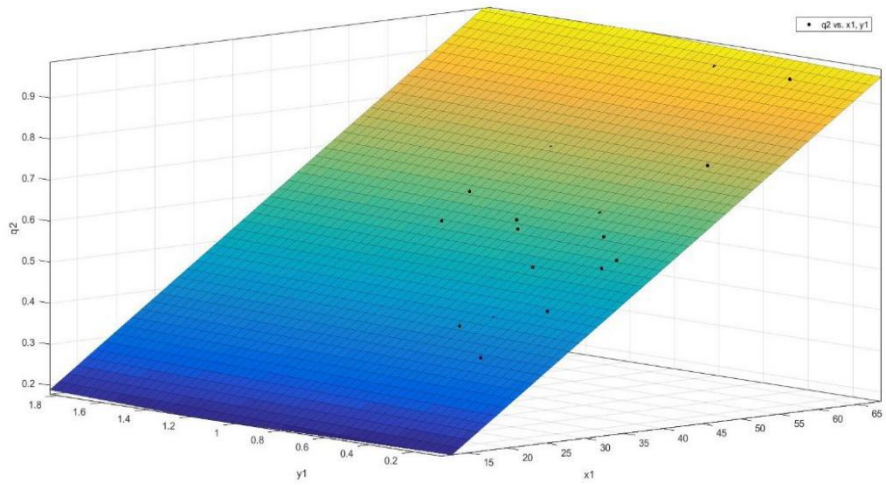
Goodness of fit:

SSE: 0.01199

R-square: 0.9851

Adjusted R-square: 0.984

RMSE: 0.02107



图表 8 完成度函数图像

4.2.3 与原方案比较结果

在本方案中，由于我们采用了附件一中成功完成任务的经验所建立模型，所以对于在附件一中已完成的任务，我们给出的方案很好的继承了下来，保持了这些任务完成度相对较高的水平；同时，我们借鉴了附件一中未完成任务的失败教训，对附件一中原本未完成的任务进行了特殊的单独优化方案，相对于已完成任务是隔离状态的，这样操作之后，可以大大提高未完成任务的完成度，而且保持了原有的已完成任务的完成度，从而在整体上提高了总完成度。虽然在少数任务上也会存在完成度较低的情况，但在全局来看，此次给出的优化方案很好地完成了任务，使得总完成度由原有的 0.625150 提高到了 0.859546。总体比原方案优化了，最终效果也在预期范围之内。

4.3 问题三模型的建立与求解

4.3.1 基于优化模型打包任务的定价

当多个任务因为位置比较集中，导致用户会争相选择的时候，将这些任务联合在一起打包发布是一个可行的方法。对于定价分配，即把不同地区的任务进行打包，理论上，在任务稀疏的地方，一个包中的任务数少一些，任务密集的地方一个包内的任务更多一些，但是一个包内的任务数应有一上限值；在定价上，一个包的定价应比原有的包内所有任务总定价低一些，这样更利于任务提供者；对于较难完成的任务，则可适当提高价格以吸引更多的人来完成。[4]

设一个任务包所覆盖区域面积为 S，任务总数为 W，原任务总价格/任务包价格=k，由实际调研及查阅资料可知，$S \le S_0$，其中S_0为平均可忍受面积范围最大值;$W \le W_0$,其中W_0为平均可忍受任务数量最大值；$k \in (0.7,1.1)$,k 具体取决于单位面积内的会员数量。接下来按照任务完成可能性的模型分步确定 S、W、和 k 的值。

对于完成度检验函数：

$$f(x,y)=-0.006036+0.0182068x-1.049737y^2 \tag{18}$$

其中 x 为一定区域内 总任务价值/任务总个数 ，y 为一定区域内 任务总数/会员总数。

例如，对于从（22.98,112.98）到（23.24,113.34）区域，总面积值为 0.0936，经 Excel 筛选、计算，得到总任务数值为 236，此区域内平均定价为 68.61 元，总会员数量为 763。计算得：

X’=68.61;

Y’=236/763=0.3093

完成度 Z’= f(X’,Y’)=0.90234

与上述计算过程相似地，整个区域内总的面积值为 2.507896，总任务值为 835，此区域内平均定价为 69.11078 元，总会员数量是 1878。计算得：

X0=69.110784

Y0=835/1878=0.444622

Z0=0.7601212

将总的图形区域分为四部分，计算每一部分的完成度，记为 Z11、Z12、Z13、Z14，并求其平均值得

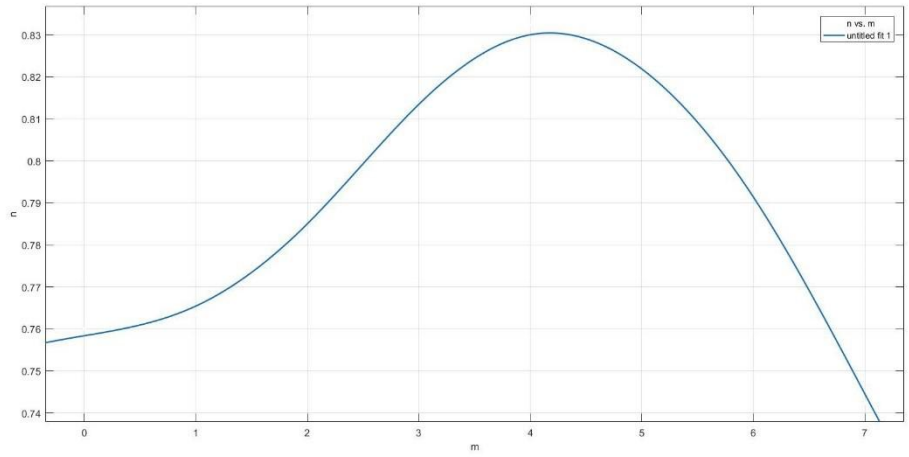
Z1=0.7634265

以此类推，求得：

Z2=0.7823439

Z3=0.8149024
Z4=0.8323485
Z5=0.8214252
Z6=0.7932458
Z7=0.7423787

总体完成度如图所示：



图表 9

依照解析式以及所计算的结果均显示，完成度在面积是 1.0542*0.9375Kmˆ2 左右的时候是较高的。如果，我们用不均匀划分的方式来分开图形，任务密集地方在一块儿，稀疏的在另一块儿。考虑到人工区分工作量巨大，计算量也很大，我们计算了 10 组数据，与均等分布相比，不均等分布的完成度更大。另外，若表示“打包后一个包的价值/未打包时任务总值”=为 u，则 u 越大，完成度越大，但增加了任务发布者的经济负担，而在完成度保持 90%以上，u 应取 0.87 以上。所以应该按照任务发布者的实际情况取 u 的值，建议 u∈（0.87,1.12）。

最终表明，打包时，一个包的地理范围大小应以约 1*1Kmˆ2 为佳，应根据会员分布的实际情况对任务进行相应的不均匀分布，打包后一个包的价值/未打包时任务总值应以任务发布者的要求选择，建议值为（0.87,1.12）。这样一来，可见打包后对最终任务完成结果（原总结果约为 76%）的优化情况是较为明显的，同时，打包之后，降低了任务发布者的定价总投入；而当一包任务在一个地理范围之内打包，也使得附近的任务接受者更为方便地获得更多的酬金。

4.4 问题四模型建立与求解

在本问中，我们利用第一二问的已经求解出的模型进行求解，首先利用第一问中已完成的点进行核回归分析求解出曲面函数：

f(x,y) = p00 + p10*x + p01*y + p20*x^2 + p11*x*y + p02*y^2 + p30*x^3 + p21*x^2*y + p12*x*y^2 + p40*x^4 + p31*x^3*y + p22*x^2*y^2

where x is normalized by mean 23.02 and std 0.2226and where y is normalized by mean 113.5 and std 0.3491，

Coefficients (with 95% confidence bounds):

p00 =	66.06	(65.75, 66.38)
p10 =	-2.308	(-2.825, -1.792)
p01 =	1.052	(0.7893, 1.314)
p20 =	1.617	(1.3, 1.933)
p11 =	1.757	(1.261, 2.252)
p02 =	2.516	(2.263, 2.769)
p30 =	0.325	(0.1041, 0.5459)
p21 =	-0.7733	(-1.121, -0.4259)
p12 =	1.484	(1.138, 1.83)
p40 =	-0.1316	(-0.1853, -0.07791)
p31 =	-0.1323	(-0.3107, 0.04603)
p22 =	0.07609	(-0.1531, 0.3052)

带入附件三 GPS 位置求出定价 1。之后将定价 1 在地图上标注。同样利用上述原理求解出第一问中未完成的点对应的核回归曲面函数，将二者结合为一张图，并利用 MATLAB 得出在（±0.0001，±0.0002）误差坐标内定价 1 和未完成任务定价相差小于等于 2 的所有点，对这些点进行单独优化，提升这些点的定价。具体提升过程是：先确定提升区间（提升后的定价/提升前的定价）为（1.0,2.0），分为 100 等份，得出 100 个对应的定价 2，对于每个定价 2，求解出对应的完成度，调整 k 值以此调整定价，取完成度高于 95%的最小的定价，记为定价 3。最终经过三次优化之后，我们得出了最终的定价，该定价记录在文件“附件三定价方案.xls”中，同时求解出每个定价对应的完成度，并求解出总完成度为 0.896425，这个完成度明显大于原有方案的 0.625150。

依照打包对总结果的优化明显程度，建议在以上定价的同时，加入打包机制，一个包的地理范围大小应以约 1*1Km^2 为佳，根据会员分布的实际情况对任务进行相应的不均匀划分，打包后一个包的价值/未打包时任务总值以任务发布者的要求选择，建议值是（0.87,1.12）。

最终，按照我们给出的定价方案，在不打包的情况下，完成度约为 XX%，明显高于题中所给的默认定价方案，较好地完成了优化任务；如果加入打包机制，我们的方案的完成度会更加优化，得到更好的效果。

五、灵敏度分析与模型检验

5.1 灵敏度分析

局部法主要分析因素对模型的局部影响（如某点）。局部法可以得到参数对输出的梯度，这一数值是许多领域研究中所需要的重要数据。局部法主要应用于数学表达式比较简单，灵敏度微分方程较易推出，不确定因素较少的系统模型中。主要包括直接求导法、有限差分法、格林函数法。这里我们采用了直接求导法。对于输入因素个数少、结构不复杂、灵敏度微分方程较易推导的系统或模型，直接求导法分析了灵敏度。[3] 假设要考虑的初值问题是

$$f(y,x)=\frac{dy}{dx}$$

$$y(0)=y^6$$

同样，y 代表 n 维输出变量，X 代表 m 维输入因素。y^0代表初值数组。

式（1）对输入因素x_j微分得到下述的灵敏度分析微分方程

$$\frac{d}{dt}\frac{\partial y}{\partial x_j}=J\frac{\partial y}{\partial x_j}+\frac{\partial f}{\partial x_j}$$

或以矩阵形式表示为

$$S=JS+F$$

式中，$J=\{\frac{\partial f_i}{\partial y_i}\}$是系统代数-微分方程右边对系统输出变量的导数。

以此使用 MATLAB 进行求解，结果为：
所有任务的“价格-经纬度”函数梯度 D1 =

[- (1593*x^3)/2500 - (387*x^2*y)/625 + (2223*x^2)/2000 + (8977*x*y^2)/5000 - (9419*x*y)/5000 + (377*x)/125 + (1131*y^2)/1000 + (1927*y)/500 - 1809/1000, (1927*x)/500 + (562*y)/125 + (1131*x*y)/500 + (8977*x^2*y)/5000 - (9419*x^2)/10000 - (129*x^3)/625 + 8197/10000, 0]

已完成任务的“价格-经纬度”函数梯度 D2 =

[- (1593*x^3)/2500 - (387*x^2*y)/625 + (2223*x^2)/2000 + (8977*x*y^2)/5000 - (9419*x*y)/5000 + (377*x)/125 + (1131*y^2)/1000 + (1927*y)/500 - 1809/1000, (1927*x)/500 + (562*y)/125 + (1131*x*y)/500 + (8977*x^2*y)/5000 - (9419*x^2)/10000 - (129*x^3)/625 + 8197/10000, 0]

完成度检验函数梯度 D3 =

[5247752812518985/288230376151711744,-(2363797581018511*y)/11258999068426624]

分析结果可知，对于我们建立的这两个主要模型，在少量数据发生突变的时候对整体模型的求解值影响程度极小，可以忽略不计；

$$\lim\frac{\frac{\Delta x}{x}\frac{\Delta y}{y}}{(\frac{\Delta z}{z})^2}=\left(\frac{z^2}{xy}\right)\left(\frac{\partial x}{\partial z}\frac{\partial y}{\partial z}\right)=(\frac{1}{k_1}\frac{1}{2}\sqrt{\frac{k_1}{z}})\left(\frac{z^2}{xy}\right)$$

当 x=68.61, y=0.3093, z=0.90234 时，原式得 1.136289，在对于摸一个具体数据，当发生微小量变动 的时候，求解结果的变化量 要比 小几个数量级，因此模型具有很好的稳定性，同时在数据变动的时候又不失灵敏度。

5.2 回归模型优度分析

1、所有任务的“价格-经纬度”函数方差等的参考值：

Goodness of fit1:

SSE: 19.12

R-square: 0.6651

Adjusted R-square: 0.6606

RMSE: 2.629

2、已完成任务的“价格-经纬度”函数方差等的参考值：

Goodness of fit2:

SSE: 10.28

R-square: 0.8323

Adjusted R-square: 0.8287

RMSE: 1.994

3、完成度检验函数方差等的参考值：

Goodness of fit3:

SSE: 0.01199

R-square: 0.9851

Adjusted R-square: 0.984

RMSE: 0.02107

下面对于上述三个模型的拟合评价参数做出解释:

一、SSE(和方差)

该统计参数计算的是拟合数据和原始数据对应点的误差的平方和。SSE 越接近于 0，说明模型选择和拟合更好，数据预测也越成功。

二、MSE(均方差)

该统计参数是预测数据和原始数据对应点误差的平方和的均值，也就是 SSE/n。

三、RMSE(均方根)

该统计参数，也叫回归系统的拟合标准差，是 MSE 的平方根。

四、R-square(确定系数)之间的误差——即点对点。而“确定系数”是所有的误差都是相对原始数据平均值而展开的——即点对全。“确定系数”是通过数据的变化来表征一个拟合的好坏。“确定系数”的正常取值范围为[0，1]，越接近 1，表明方程的变量对 y 的解释能力越强。

根据上述描述，可以看出上面三个拟合方程符合正常计算对于四个参数 SSE、MSE、RMSE、R-square 的要求。因此得到结论上述模型符合实际要求。

5.3 模型检验

运用完成度检验模型，将数据拟合，到 0.98 的方差，拟合的特别好，由分析问题二中的模型，可以得到总完成度由原有的 0.625150 提高到了 0.859546，说明此模型优于旧定价模型。

六、模型评价

6.1 优点

1、运用核回归对散乱点进行三维连续曲面拟合，不对函数形式做任何假定，可以降低因模型结构带来的误差；

2、建立了一件任务的完成度，来衡量一件任务完成的可能性，该检验模型充分考虑了任务和会员所在地的位置及其重合程度，一定区域内总的任务数量、总的会员数量及其比例等因素，并采用一系列数据来检验这个检验模型，最终确定了模型的可靠性。

3、对模型进行了稳定性、健壮性、灵敏度的分析，更为科学、合理地验证了模型的正确性和合理性；对所有模型进行了大量数据测试，保证并证明了所有模型在处理大量数据的时候的结果可靠性。

4、通过利用 MATLAB、C++、Excel 函数等一系列工具来求解、测试和修正模型，在降低工作量的同时提高了模型的精准度，同样在做每一步计算的过程中，我们都在保证正确的前提下采用最大精度，之后在实际书写论文的时候我们均保留了六位有效数字（某些特殊情况除外）。以上这些都提高了模型的精准度。

6.2 不足

1.在大方向上采用了主成分分析法，忽略了次要因素的影响。虽然这些次要因素影响程度较小，但是仍然会对模型造成微小扰动，但在时间和成本等综合考量后仍然选择忽略这些因素，如果有更好的条件，此模型会更为精确。

2.在整个过程中有一些操作涉及到人为选择，虽然这些操作数量非常小并且我们选择了三位成员分别操作最后取平均值的方式，但仍会造成一些微小的误差。

参考文献

[1]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.
[2]赵亮,赵春霞,张二华. 核回归方法的散点拟合曲面重构[J]. 计算机研究与发展,2009,46(09):1446-1455. [2017-09-17].
[3]韩林山,李向阳,严大考. 浅析灵敏度分析的几种数学方法[J]. 中国水运(下半月),2008,(04):177-178. [2017-09-17].
[4]张媛. 大众参与众包的行为影响因素研究[D].东北财经大学,2011.
[5]姜启源,谢金星,叶俊. 数学模型[D].高等教育出版社,2011.

附录一

求解完成度模型使用的 matlab 核心代码：

```
a=22.64;%所选方块纬度（低）
b=22.76;%所选方块纬度（高）
c=114.24;%所选方块经度（低）
d=114.42;%所选方块经度（高）
>> tasklength=length(find(a2(:,1)>=a&a2(:,1)<=b&a2(:,2)>=c&a2(:,2)<=d));
>> finishlength=length(find(a2(:,1)>=a&a2(:,1)<=b&a2(:,2)>=c&a2(:,2)<=d&a2(:,4)==1));
>> peoplelength=length(find(a2_1(:,1)>=a&a2_1(:,1)<=b&a2_1(:,2)>=c&a2_1(:,2)<=d));
>> sum=0;
>> for i=1:1:826
if(a2(i,1)>=a&a2(i,1)<=b&a2(i,2)>=c&a2(i,2)<=d&a2(i,4)==1)sum=sum+a2(i,3);
end
end
>> q1=finishlength/tasklength;
>> x=sum/tasklength;
>> y=tasklength/peoplelength;
>> %至此对于该方块三个量——完成度、总价/任务总数、任务数/人数求解完成
```

附录二

由于附件二中的会员位置信息是连在一起的，matlab 无法直接分离出来，因此我们编写了提取附件二中会员位置信息的 c++程序：

```
#include <fstream>
#include <string>
#include <iostream>
```

```
#include <iomanip>
using namespace std;

int main()
{
	ifstream in("1.txt");
	ofstream out("2.txt");
	string filename;
	string line;
	double a,b;

	if(in) // 有该文件
	{
		while (getline (in, line)) // line 中不包括每行的换行符
		{
			in>>a>>b;
			out << setprecision(15) << b << endl; // 输入到 2.txt 中
		}
	}
	else // 没有该文件
	{
		cout <<"no such file" << endl;
	}

	return 0;
}
```