“拍照赚钱”的任务定价

摘要

随着劳务众包平台的兴起，如何对任务进行合理定价成为平台的重要工作。本文依据题目所给数据研究现有定价规律，并建立模型进行优化。

对于问题一，首先我们通过数据预处理将 **13** 个异常会员值进行剔除。之后，运用**墨卡托投影法**，将所有数据的经纬度转换成平面坐标。然后，我们用 **k-means** 聚类方法对任务点作了聚类分析，计算了偏僻程度，会员密度，任务密度和任务标价的 **Pearson 相关系数**，并构建**多元线性回归模型**来拟合任务标价规律。最后，我们分析了任务未完成的 3 个原因。

对于问题二，我们结合当地的经济发展情况和任务、会员之间的距离，定义了任务对会员**吸引力**这个概念。之后，我们建立**多目标规划模型**，求解出了新的任务标价和任务完成情况。新方案比原方案成本降低 **13.07%**，完成率提升 **29.71%**。

对于问题三，我们首先求出任务之间的距离矩阵，然后通过**层次聚类法**对相邻任务进行打包。之后，对第二问的多目标优化模型参数进行修正，求得新的方案。新方案比问题二的方案成本降低 **8.7%**，完成率提升 **6.13%**。

对于第四问，我们先用层次聚类法对任务进行打包。之后，通过 **BP 神经网络**模型，以经纬度作为输入量，任务标价、完成情况作为输出量，求得任务定价方案，完成率为 **85.97%**

最后，我们对模型进行了优缺点分析和推广。

关键字： 墨卡托投影法　层次聚类法　多元线性回归模型　多目标优化模型

目录

一、 问题重述

1.1 背景资料

随着现代科技的飞速跃进，我们的世界以前所未有的方式互联起来，全球生产力因此得到了前所未有的提升，世界各地的联系也日益紧密。然而，这种在线连接的广泛性和便捷性，也极大地暴露了我们个人和集体的网络安全脆弱性，使得网络犯罪问题日益严重。网络犯罪的查处之所以困难重重，其核心在于其跨国界的特性，导致调查和起诉工作在管辖权上面临诸多复杂和棘手的挑战。许多机构，如投资公司，为了维护自身的声誉和客户信任，往往选择对黑客攻击事件秘密处理，宁愿私下支付赎金以解决问题，也不愿让客户和潜在客户知道其网络安全存在漏洞。

为了有效应对网络犯罪带来的日益严峻的挑战和风险，各国政府纷纷制定并公布了国家网络安全政策，旨在构建更加安全、稳定的网络环境，保护公民和企业的合法权益。在此过程中，国际电信联盟（ITU）作为联合国的专门机构，在信息和通信技术领域发挥着至关重要的引领作用。ITU 不仅致力于制定国际标准，规范网络安全行为，确保各国在网络安全方面的合作与协调；还积极促进国际合作，加强各国在打击网络犯罪方面的协同作战能力，共同应对网络犯罪的全球化挑战。同时，ITU 还开发了一系列科学、有效的评估工具，帮助全球和各国准确衡量网络安全状况，为制定更加精准、有效的网络安全策略提供有力支撑和科学依据。

GPT：With the rapid advancement of modern technology, our world is interconnected in unprecedented ways, leading to an unparalleled boost in global productivity and increasingly close ties among different parts of the world. However, the widespread and convenient nature of this online connectivity has also greatly exposed our individual and collective vulnerabilities in cybersecurity, making the issue of cybercrime increasingly severe. The core reason for the difficulty in investigating and prosecuting cybercrime lies in its cross-border nature, which poses numerous complex and challenging jurisdictional issues for investigation and prosecution efforts. Many organizations, such as investment firms, often choose to handle hacker attacks secretly, preferring to pay ransoms privately to resolve the issues rather than letting clients and potential clients know about their cybersecurity vulnerabilities.

In order to effectively address the increasingly severe challenges and risks posed by cybercrime, governments around the world have formulated and published national cybersecurity policies aimed at building a safer and more stable cyber environment and protecting the legitimate rights and interests of citizens and enterprises. In this process, the International Telecommunication Union (ITU), as a specialized agency of the United Nations, plays a crucial leading role in the field of information and communication technology. ITU is not only committed to developing international standards, regulating cybersecurity behavior, and ensuring cooperation

and coordination among countries in cybersecurity; it also actively promotes international cooperation, strengthens the collaborative combat capabilities of countries in fighting cybercrime, and jointly addresses the global challenge of cybercrime. At the same time, ITU has developed a series of scientific and effective assessment tools to help the global community and individual countries accurately measure their cybersecurity status, providing strong support and scientific basis for formulating more precise and effective cybersecurity strategies.

二、 问题分析

2.1 问题一分析

为了探究网络犯罪在全球范围内的分布，我们首先需要收集各国网络犯罪的相关数据，包括犯罪数量、类型以及目标国家等信息。这些数据可以来源于国际组织、国家安全机构以及网络安全公司发布的报告和统计数据。随后，我们对数据进行预处理，清洗掉重复、不完整或异常的值，并进行分类和整理。接着，我们运用描述性统计方法和地图可视化工具，对网络犯罪的全球分布进行描述和展示。通过聚类分析等方法，我们可以识别出网络犯罪的高发区域和模式，进而分析哪些国家不成比例地成为网络犯罪的高目标，以及网络犯罪在不同地区的成功、被阻止、被报告和被起诉的情况，总结出网络犯罪分布的模式和特征。

GPT:To explore the distribution of cybercrime globally, we first need to collect relevant data on cybercrime from various countries, including information on the number of crimes, types of crimes, and target countries. This data can be sourced from reports and statistical data released by international organizations, national security agencies, and cybersecurity companies. Subsequently, we preprocess the data to remove duplicate, incomplete, or abnormal values, and then classify and organize it. Next, we use descriptive statistical methods and map visualization tools to describe and display the global distribution of cybercrime. Through methods such as cluster analysis, we can identify high-incidence areas and patterns of cybercrime, and further analyze which countries disproportionately become high targets of cybercrime, as well as the situations where cybercrime is successful, prevented, reported, and prosecuted in different regions. Ultimately, we summarize the patterns and characteristics of the distribution of cybercrime.

2.2 问题二分析

为了比较各国发布的国家安全政策与网络犯罪分布的关系，我们需要收集国家安全政策的文本和实施效果数据，并与网络犯罪分布数据进行匹配。首先，对政策文本进行编码和分类，提取关键信息和特征。然后，运用相关性分析、回归分析等方法，探究国

家安全政策与网络犯罪分布之间的相关性。通过比较不同政策在预防、起诉和缓解网络犯罪方面的效果，并考虑政策的采用时间，我们可以识别出哪些政策或法律条款在应对网络犯罪方面特别有效或特别无效。最终，总结出政策效果与网络犯罪分布之间的模式和规律，为政策制定和完善提供数据支持。

GPT: To compare the relationship between national cybersecurity policies released by various countries and the distribution of cybercrime, we need to collect the text of national cybersecurity policies and data on their implementation effects, and match them with cybercrime distribution data. Firstly, we encode and categorize the policy texts to extract key information and features. Then, we use methods such as correlation analysis and regression analysis to explore the correlation between national cybersecurity policies and the distribution of cybercrime. By comparing the effectiveness of different policies in preventing, prosecuting, and mitigating cybercrime, and considering the adoption time of the policies, we can identify which policies or legal provisions are particularly effective or ineffective in addressing cybercrime. Finally, we summarize the patterns and laws between policy effects and the distribution of cybercrime, providing data support for policy formulation and improvement.

2.3 问题三分析

为了探究国家人口统计数据与网络犯罪分布的相关性，我们需要收集各国的人口统计数据，如互联网接入率、人均财富、教育水平等，并与网络犯罪分布数据进行匹配。接着，运用相关性分析、回归分析等方法，分析不同人口统计指标对网络犯罪分布的影响程度和方向。通过这一分析，我们可以讨论人口统计数据如何支持或混淆网络犯罪分布的理论，提出可能的解释和假设。例如，互联网接入率高的国家可能网络犯罪数量也相对较多，而教育水平高的国家可能网络犯罪率相对较低。这些分析有助于我们更深入地理解网络犯罪的分布规律，为制定针对性的网络安全政策提供科学依据。

GPT:To explore the correlation between national demographic statistics and the distribution of cybercrime, we need to collect demographic statistics from various countries, such as internet access rates, per capita wealth, education levels, etc., and match them with cybercrime distribution data. Then, we use methods such as correlation analysis and regression analysis to analyze the extent and direction of the influence of different demographic indicators on the distribution of cybercrime. Through this analysis, we can discuss how demographic statistics support or confuse theories about the distribution of cybercrime and propose possible explanations and hypotheses. For example, countries with high internet access rates may also have a relatively high number of cybercrimes, while countries with high education levels may have a relatively low cybercrime rate. These analyses help us gain a deeper understanding of the distribution patterns of cybercrime and provide a scientific basis for formulating targeted cy-

bersecurity policies.

三、 符号说明

符号	意义
L_n	经度
L_a	纬度
d_i	任务 i 的偏僻程度
ρ_j	会员密度
ρ_i	任务密度
L_a	纬度
a_{ij}	任务 i 对会员 j 的吸引力
a_i	吸引力阈值
p_i	任务 i 的标价

四、 模型的建立与求解

4.1 问题一模型的建立及求解

4.1.1 模型的准备：数据可视化与处理

根据 VERIS 社区数据库以及其他网站的数据，我们通过遍历法，统计出各个国家遭受到网络犯罪的次数与发起网络犯罪的次数；各个国家的网络犯罪成功率与挫败率；各个国家网络犯罪的举报次数与被起诉次数。

1. 网络犯罪

因为无论是各个国家遭受到网络犯罪的次数或发起网络犯罪的次数都不能准确表达该国家的网络犯罪次数，所以我们定义每个国家网络犯罪次数 = 每个国家被网络犯罪攻击的次数 + 每个国家发起网络犯罪的次数。即可得出各个国家网络犯罪的分布情况。

我们将在易受攻击国家分布地图上标记出来，如下图所示：

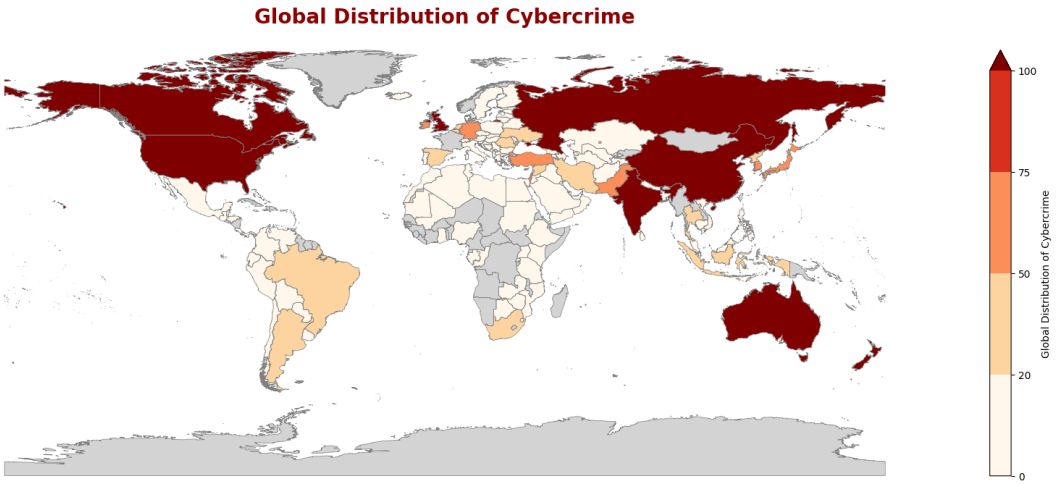

图 **1**　网络犯罪分布图

从图中可以发现，网络犯罪主要集中在美国、加拿大、英国、俄罗斯、中国、印度、澳大利亚、新西兰八个国家。

2. 易受攻击国家分布

易受攻击国家为受攻击数加上攻击数的排名靠前的国家，我们将在易受攻击国家分布地图上标记出来，如下图所示：

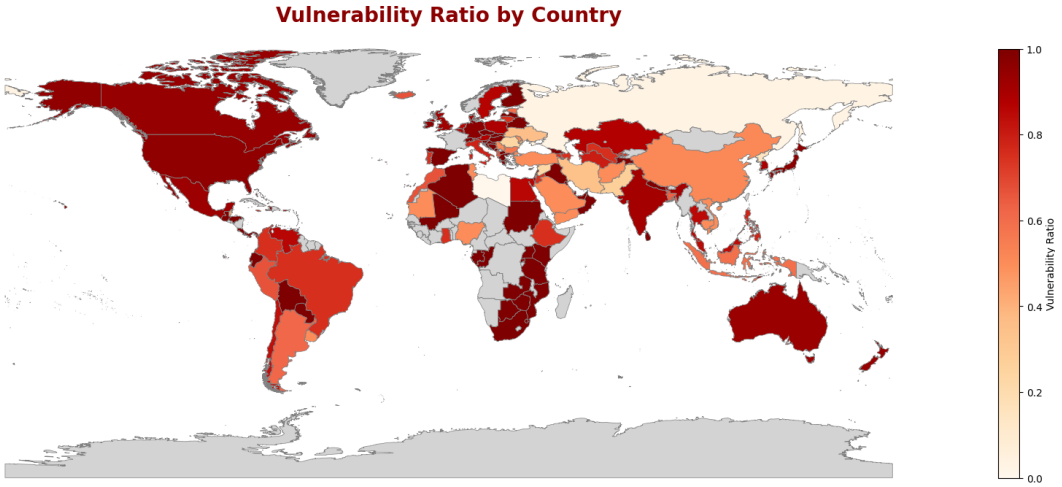

图 **2**　易受攻击国家分布图

从图中我们可以看出，网络犯罪分布情况和易受攻击国家分布情况有相似之处，因此我们可以推测，网络犯罪分布情况和易受攻击国家分布情况存在一定关系。

3. 网络犯罪成功率

我们将在网络犯罪成功率分布地图上标记出来，如下图所示：

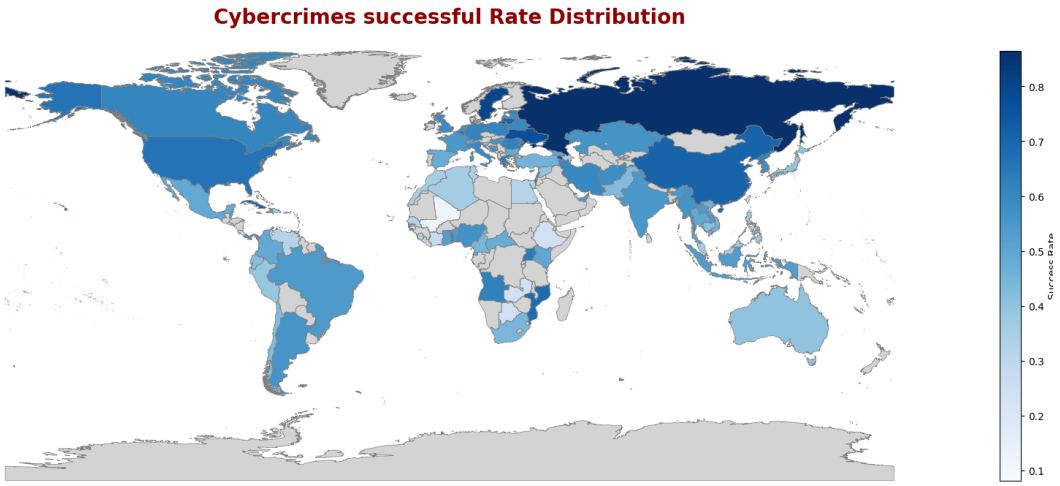

图 **3**　网络犯罪成功率分布图

从图中我们可以看出，网络犯罪成功率分布情况和网络犯罪分布情况有相似之处，因此我们可以推测，网络犯罪成功率分布情况和网络犯罪分布情况存在一定关系。

4. 网络犯罪挫败率

我们将在网络犯罪挫败率分布地图上标记出来，如下图所示：

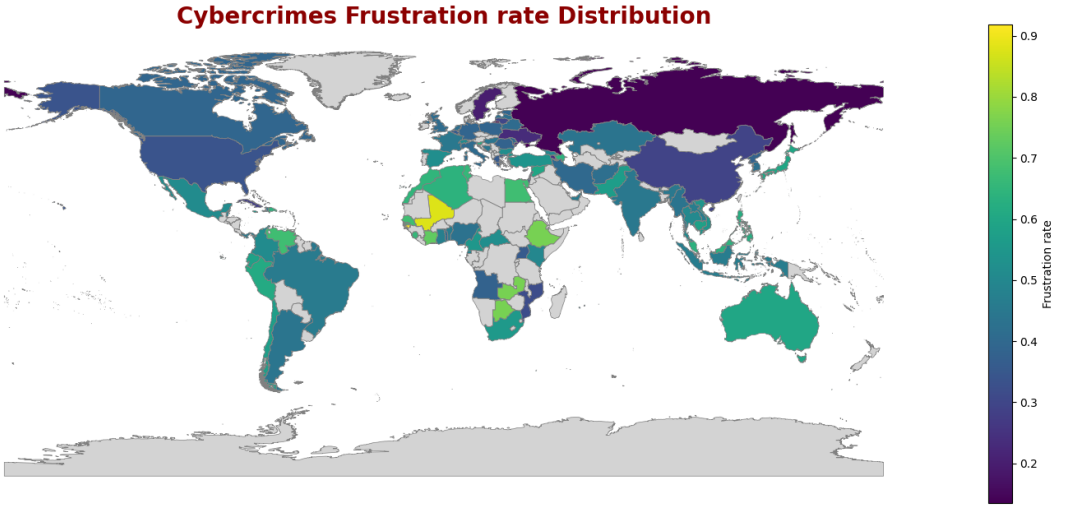

图 **4**　网络犯罪挫败率率分布图

从图中我们可以看出，网络犯罪挫败率分布情况和网络犯罪成功率分布情况有相似之处，因此我们可以推测，网络犯罪挫败率分布情况和网络犯罪成功率分布情况存在一定关系。

5. 网络犯罪举报分布

我们将在网络犯罪举报分布地图上标记出来，如下图所示：

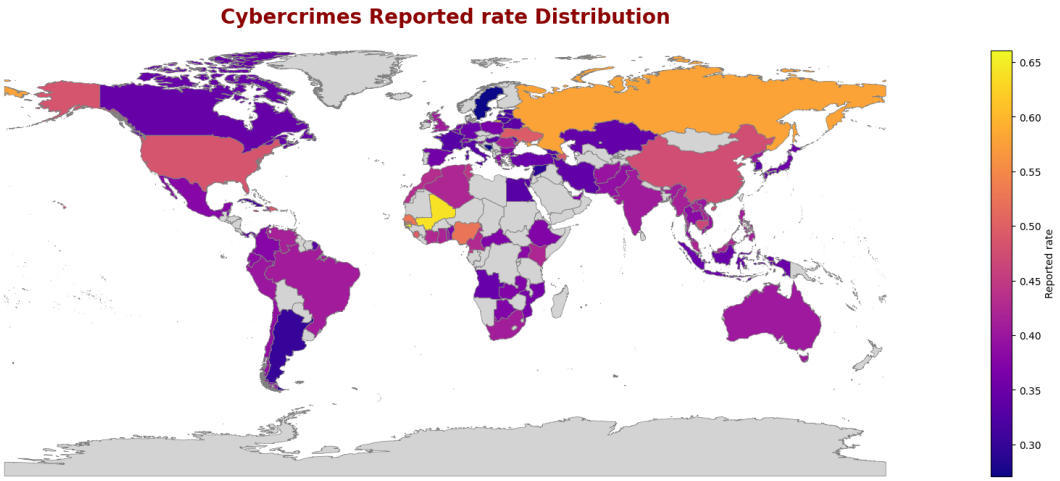

图 **5**　网络犯罪举报分布图

从图中我们可以看出，网络犯罪举报分布情况和网络犯罪成功率分布情况有相似之处，因此我们可以推测，网络犯罪举报分布情况和网络犯罪成功率分布情况存在一定关系。

6. 网络犯罪被起诉分布

我们将在网络犯罪被起诉分布地图上标记出来，如下图所示：

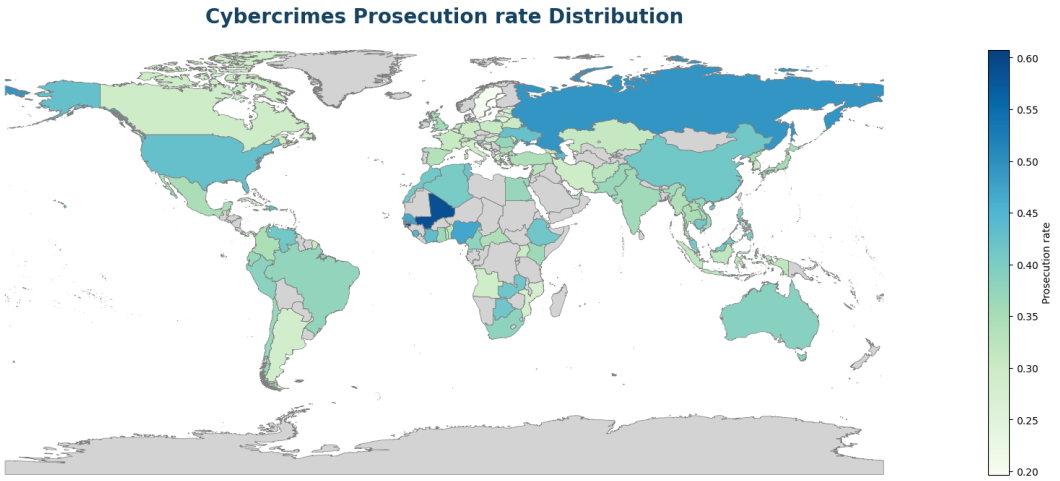

图 **6**　网络犯罪被起诉分布图

从图中我们可以看出，网络犯罪挫败率分布情况和网络犯罪举报分布情况有相似之处，因此我们可以推测，网络犯罪挫败率分布情况和网络犯罪举报分布情况存在一定关系。

4.1.2 得到的网络犯罪模式

4.2 问题二模型的建立及求解

4.2.1 模型的准备

通过查询相关文献和分析一些国家安全政策，我们将国家安全政策分为政治、经济、社会、外交、网络安全政策五个大类，25 个小类，其中政治安全政策政治因素用世界银行的世界治理指标 (WGI) 的五个维度，设言论与问责, 政治稳定与无暴力，政府效能，监督质量，腐败控制来衡量；社会因素用人口，劳动力人口，教育程度，社会福利, 卫生状况来衡量；经济因素用 GDP 增长，就业率，物价稳定度，基尼系数，贫困率来衡量；技术因素用基础设施，互联网产业发展情况，互联网用户数，国际宽带数量，固定宽带用户数来衡量；网络安全因素用全球网络安全指数 (GCL) 的五个维度，法律措施，技术措施，组织措施，能力发展措施，合作措施。总计 25 个因素，最后从中挑选出了 10 个联系最紧密的小类。

选取 90 个政策数据比较齐全的国家（其中 30 个中高收入国家，30 个中低收入国家，30 个低收入国家），分别对每个国家的这 10 个政策进行打分，记得分为 $x_1, x_2 \cdots x_10$，打分方式选择模糊综合评价。

1. 因素集 U

对一个国家某项政策进行打分，需要从多个方面进行综合评判，记因素集为

$$U = \{u_1, u_2, u_3\} \tag{1}$$

其中 u_1 表示政府执行情况，u_2 表示社会普及度，u_3 表示群众满意度。

2. 评语集 V

每个指标的评价值不同，因此会形成不同等级，记评语集为

$$V = \{5points, 4points, 3points, 2points, 1point, 0points\} \tag{2}$$

3. 各因素权重 A

这里认为三个因素同等重要，记权重向量为：

$$A = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \tag{3}$$

4. 模糊综合判断矩阵 R

对指标 u_1 由政策方面的专家打分，u_2 由当地媒体打分，u_3 由当地群众打分，记模糊综合判断矩阵为

$$R =: \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{13} & r_{16} \\ r_{21} & r_{22} & r_{23} & r_{24} & r_{25} & r_{26} \\ r_{31} & r_{32} & r_{33} & r_{34} & r_{35} & r_{36} \end{bmatrix} \tag{4}$$

5. 评判结果 $Score$

记 $C = \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \end{bmatrix}^{\top}$

$$\text{Score} = A \cdot R \cdot C \tag{5}$$

4.2.2 网络犯罪的模型

有了上面的一些定义之后，我们可以建网络犯罪模型。

1. 变量的确定设变量 Y_1 为网络犯罪挫败率，Y_2 为网络犯罪举报率，Y_3 为网络犯罪起诉率。

分别以 Y_i 为因变量，$x_1 \sim x_n$ 为自变量进行线性回归。

将分为前文提到三类国家来构建三种犯罪模型。

2. 回归方程的性质

• 总体显著性检验

判断 P 值是否小于 0.05，表明线性回归模型总体上有统计学意义。

• 偏回归系数显著性检验

某个自变量的 t 检验概率 P 值 <0.05，其对因变量有显著影响；删去没有显著影响的变量。

• 共线性

一般认为 VIF 大于 10，说明有较严重共线性问题；删去变量。

注：三个方程中删去的变量可能不同。

综上，得到回归方程。

3. 模型的求解根据回归方程，我们可以预测出变量对三种不同收入水平的国家的网络犯罪的影响。

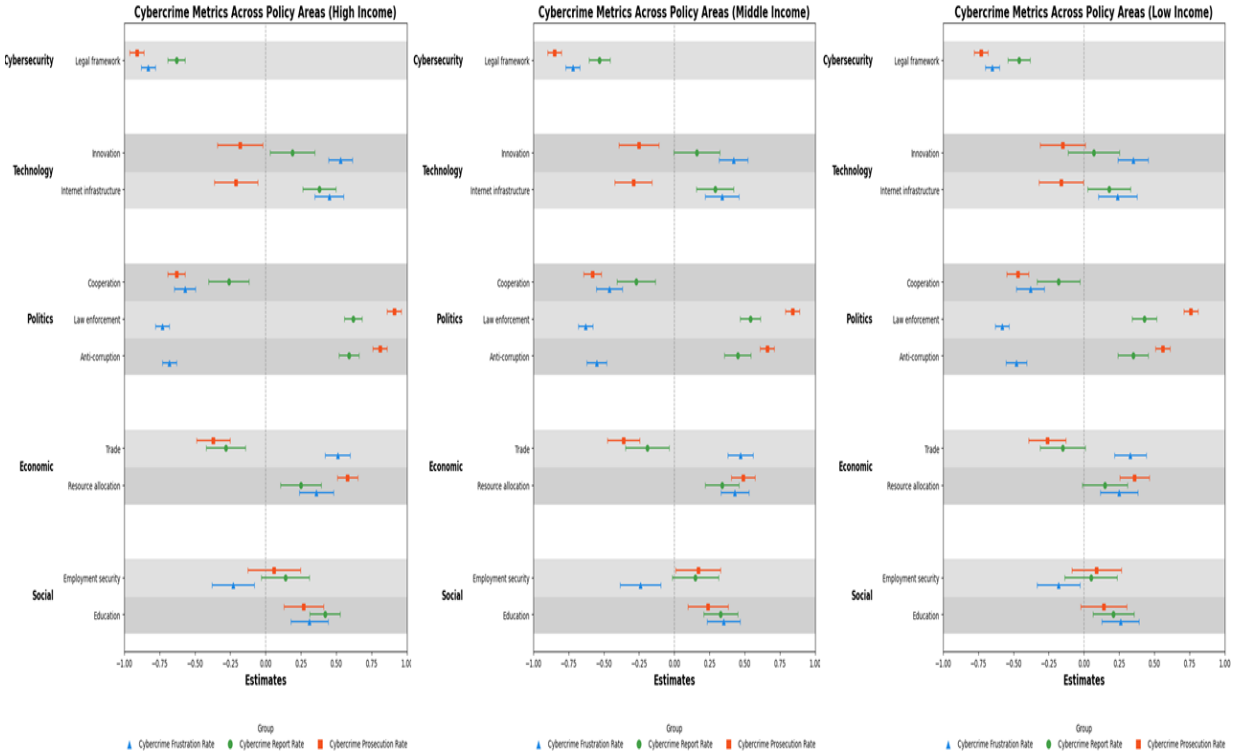

图 7　**Cybercrime Metrics Across Policy Areas**

4. 模型的检验

我们可以通过某项政策颁布前后的数据变化来验证我们的模型。

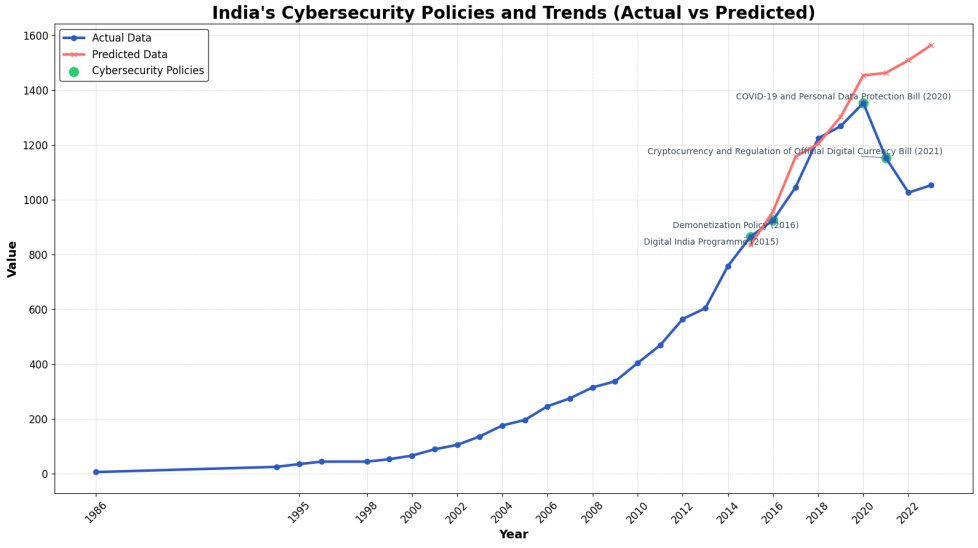

图 8　**India's Cybersecurity Policies and Trends (Actual vs Predicted)**

查阅资料得知, 印度在 2015 年准备开始实施数字印度计划，且在 2016 年开始非货币化政策，且在 2021 年颁布加密货币与官方数字货币监管法案，我们通过上图可以发现这期间模型的预测数据比实际数据多，可以推测出政策的颁布对网络犯罪有效。

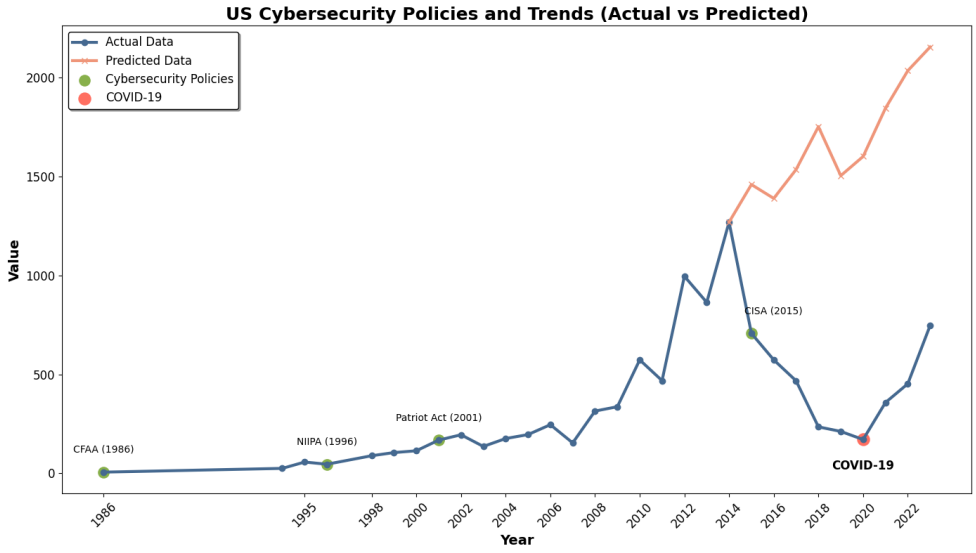

图 9　**US Cybersecurity Policies and Trends (Actual vs Predicted)**

查阅资料得知, 美国在 2015 年 CISA 发布政策，我们通过上图可以发现这期间模型的预测数据比实际数据多，可以推测出政策的颁布对网络犯罪有效。

4.3 问题三模型的建立及求解

4.3.1 网络犯罪分布

1. 建立网络犯罪的关系网络图

为制定强有力的国家安全政策，我们需要更深入地了解网络犯罪的分布模式。我们构建一个概念框架，包括政治，社会，经济技术，网络安全因素。通过查阅相关文献，我们可初步得出它们之间的关系网络图。

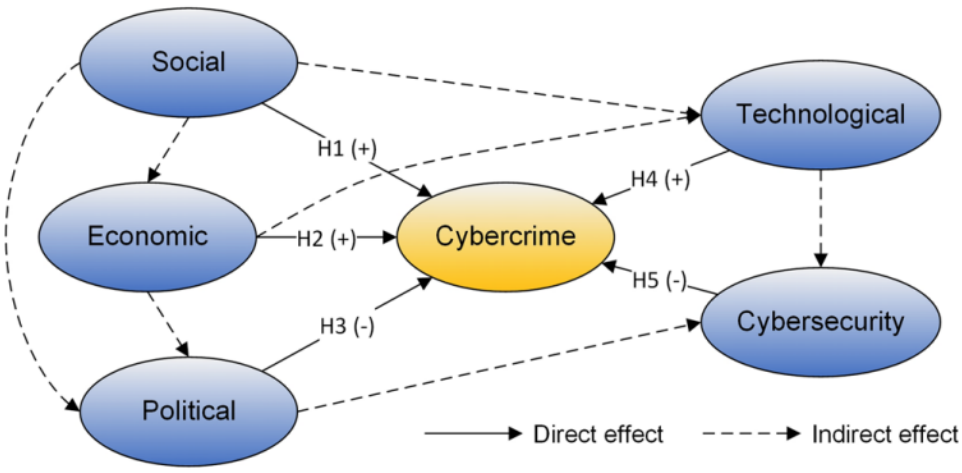

图 10　**The conceptual framework for analysing driving forces of cybercrime.**

概念框架中的五个因素作为潜在变量，我们无法直接观测他们的数值，我们将用一系列测量变量去描述它们。其中政治因素用世界银行的世界治理指标 (WGI) 的五个维度，设言论与问责 ∂_1, 政治稳定与无暴力 ∂_2，政府效能 ∂_3，监督质量 ∂_4，腐败控制 ∂_5 来衡量；社会因素用人口 ∂_6，劳动力人口 ∂_7，教育程度 δ_8，社会福利 ∂_9, 卫生状况 ∂_{10} 来衡量；经济因素用 GDP 增长 ∂_{11}，就业率 ∂_{12}，物价稳定度 ∂_{13}，基尼系数 ∂_{14}，贫困率 ∂_{15} 来衡量；技术因素用基础设施 ∂_{16}，互联网产业发展情况 δ_{17}，互联网用户数 ∂_{18}，国际宽带数量 ∂_{19}，固定宽带用户数来衡量 ∂_{20}；网络安全因素用全球网络安全指数 (GCL) 的五个维度，法律措施 ∂_{21}，技术措施 ∂_{22}，组织措施 ∂_{23}，能力发展措施 ∂_{24}，合作措施 ∂_{25}。所有数据均需要作对数转换并作归一化处理。

对网络犯罪需要用网络犯罪数量作归一化的数据（设为 β）表示；以 β 为因变量，$\partial_1 \sim \partial_{25}$ 为自变量进行线性回归分析。求解步骤与第二问一致，可以得到：

$$\beta = a_1\partial_1 + a_2\partial_2 + \quad + \cdots + a_{25}\partial_{25} + a_{26} \tag{6}$$

若变量因无显著影响或共线性而被删去，则令其系数为 0。设 ∂_i 对 β 的影响率为 b_i, 令

$$b_i = \frac{|a_i|}{\sum_{i=1}^{25}|a_j|} i = 1, \cdots, 25 \tag{7}$$

我们用线性回归估计了各测量变量对犯罪分布的影响, 结果如下：

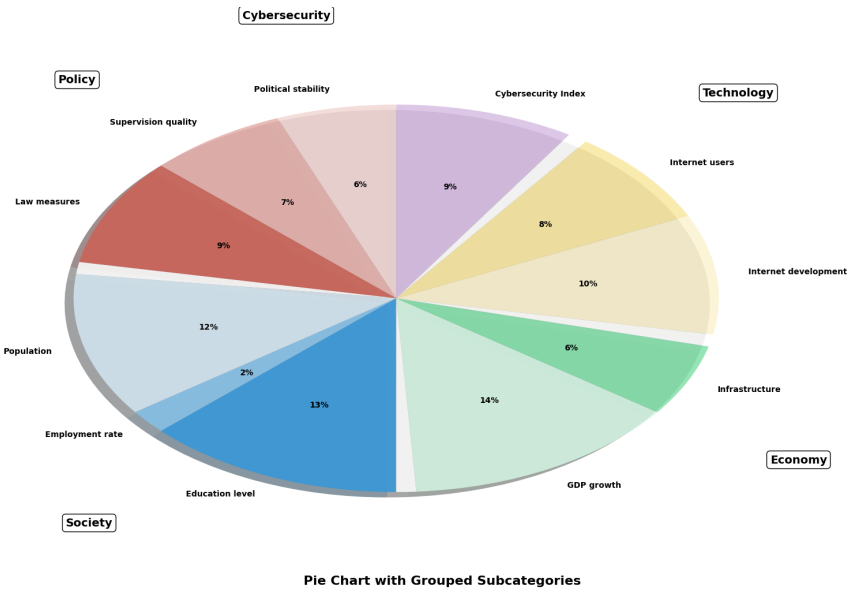

图 11　**The influence of each measurement variable on crime distribution**

2. 结构方程模型 (SEM)

现在，用结构方程模型 (SEM) 估计潜在变量之间的相互作用以及它们对犯罪分布的直接和间接影响。

結构方程模型（SEM）是一种综合统计技术，可帮助力研究人员分析潜在变量和观测变量之间的复杂关系，它结合了因子分析和多元回归的元素，能够在单个模型中检查直接和间接关系，在本文中 AMOS 软件进行模型分析。

結构方程模型（SEM）流程：

Step1：模型构建

Step2：模型拟合度检验

●CFI(Comparative Fit Index) 反映了模型相对于独立模型的改进程度，CFI 的值越接近于 1，表示模型的拟合效果更好。

●RMESA(Root Mean Square Error of Approximation) 衡量模型自由度的拟合误差，通常认为 RMESA<0.05 表示良好拟合，小于 0.08 表示可接受拟合。

●SRMR(Standardized Boot Mean Square Residual) 衡量模型预测值与观察值之间的残差，一般认为 SRMR 小于 0.08 表示良好拟合。Step3：路径分析和效应检验

通过路径系数反应各潜在变量对网络犯罪分布的直接或间接的影响。

Step4：模型调整

当测量关系不好时，将相关测量项删除后再次尝试进行分析。

4.4 提出的理论

我们的理论主要阐述了如何制定强有力的国家网络安全政策，包括五个核心原则：

首先，强调通过比较找出并优先制定有效政策。并非所有网络安全政策对网络犯罪的防治效果相同，因此应识别并优先实施效果最佳的政策，如网络安全检测。

其次，政策制定需顺应国家发展水平。不同发展水平的国家对同一政策的反应各异，如人才培养政策对中高收入国家效果显著，但对低收入国家可能产生负面影响。因此，政策制定需结合国家实际情况。

第三，全面性至关重要。网络安全政策应覆盖所有领域，以应对复杂多变的挑战。网络安全指数高的国家，其政策往往完善且覆盖全面。因此，制定政策时应逐步深入，确保全面覆盖。

第四，协调性要求政策制定时注重跨部门合作。网络犯罪与政治、经济、社会紧密相连，要求政策制定者加强信息共享和资源优化配置，实现各领域安全相互协调和促进。

第五，可持续性强调政策应着眼未来，支持长期安全目标。如网络安全项目研发和人才培养，虽短期内效果不明显，但长期能显著提升网络安全水平。因此，制定政策时不能盲目追求短期利益。

此外，人口统计数据对网络犯罪分布有显著影响，涉及社会、经济、技术等多方面因素。这些因素不仅影响网络犯罪分布，还塑造犯罪发生的宏观环境。政策制定者应充分考虑这些因素，制定综合性政策，如经济扶持、产业升级、提升居民生活水平和就业机会等，从根本上减少网络犯罪诱因。同时，社会层面和技术层面的因素也不容忽视。

政策制定者应结合人口老龄化、城市化等社会特征，制定针对性预防措施，并密切关注技术发展趋势，更新和完善法律法规，提升技术防范能力。

总之，面对网络犯罪，政策制定者需具备全局视野和长远眼光，综合考虑人口统计数据反映的多方面因素，制定全面且具有可持续性的政策，以有效提升国家网络安全水平。

五、 模型的评价

5.1 模型的优点

1. 进行了数据预处理，数据的准确率较高。
2. 将经纬度转换为平面坐标进行计算，进一步提升了计算的准确性。
3. 四个问题的模型联系紧密，层层递进。

5.2 模型的缺点

1. 对任务和会员之间的距离仅考虑直线距离，未考虑其它因素造成距离上的改变。
2. 吸引力阈值自主确定，具有一定的主观性。
3.BP 神经网络模型可解释性较差。

六、 模型的推广

本文模型的应用背景是基于智能手机和移动互联网的劳务众包平台[1]，在相似背景下的应用众多，如外卖应用，滴滴打车，快递跑腿服务平台等都涉及到商品定价与地理位置、会员积极程度的关系[2]。

参考文献

[1] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. arXiv e-prints, 2018.

[2] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In IEEE Conference on Computer Vision & Pattern Recognition, pages 6517–6525, 2017.

附录 A　墨卡托投影法

```
function [x,y]=ll_xy(lng, lat)
earthRad = 6378137.0;
x = ((lng .* pi) ./ 180) .* earthRad;
a = (lat .* pi) ./ 180;
y = (earthRad ./ 2) .* log((1.0 + sin(a)) ./ (1.0 - sin(a)));
end

tic
format long g
[x_p ,y_p] = ll_xy(x,y);
x_p = x_p - mean(x_p);
y_p = y_p - mean(y_p);
toc
```

附录 B　Kmeans-聚类算法

```
opts = statset('Display','final');
%调用 Kmeans 函数
%X N*P 的数据矩阵
%Idx N*1 的向量,存储的是每个点的聚类标号
%Ctrs K*P 的矩阵,存储的是 K 个聚类质心位置
%SumD 1*K 的和向量,存储的是类间所有点与该类质心点距离之和
%D N*K 的矩阵，存储的是每个点与所有质心的距离;
[Idx,Ctrs,SumD,D] = kmeans(X,4,'Replicates',2,'Options',opts);
%画出聚类为 1 的点。X(Idx==1,1),为第一类的样本的第一个坐标; X(Idx==1,2为第二类的样本的第二个坐标
plot(X(Idx==1,1),X(Idx==1,2),'r.','MarkerSize',14)
hold on
plot(X(Idx==2,1),X(Idx==2,2),'b.','MarkerSize',14)
hold on
plot(X(Idx==3,1),X(Idx==3,2),'g.','MarkerSize',14)
hold on
plot(X(Idx==4,1),X(Idx==4,2),'y.','MarkerSize',14)
%绘出聚类中心点,kx 表示是圆形
plot(Ctrs(:,1),Ctrs(:,2),'kx','MarkerSize',14,'LineWidth',4)
%legend('Cluster 1','Cluster 2','Cluster3','Centroids','Location','NW')
Ctrs
SumD
```

附录 C　层次聚类法

```
import pandas as pd
import seaborn as sns # 用于绘制热图的工具包
from scipy.cluster import hierarchy # 用于进行层次聚类，话层次聚类图的工具包
from scipy import cluster
import matplotlib.pyplot as plt
from sklearn import decomposition as skldec # 用于主成分分析降维的包

from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
from matplotlib import pyplot as plt

df = pd.read_excel("tempdata.xlsx", index_col=0, header=None)
    #index_col=0指定数据中第一列是类别名称，PS：计算机程序一般从整数0开始计数，所以0就代表第一列
# df = df.T  #python默认每行是一个样本，如果数据每列是一个样本的话，转置一下即可

X = df.index
# print (X)
# method是指计算类间距离的方法,比较常用的有3种:
# single:最近邻,把类与类间距离最近的作为类间距
# average:平均距离,类与类间所有pairs距离的平均
# complete:最远邻,把类与类间距离最远的作为类间距
Z = linkage(X, 'average')
f = fcluster(Z, 4, 'distance')
fig = plt.figure()
dn = dendrogram(Z)
plt.show()
```