|  |  |  |
| --- | --- | --- |
| 选题 | **2024年第十四届APMCM**  **亚太地区大学生数学建模竞赛（中文赛项）** | 参赛编号 |
| B | apmcm24102596 |

标题（此处换成论文的标题）

摘要

(说明：以下开始写摘要，正文从下一页开始。摘要及正文格式基本要求是宋体，小四号，单倍行距，没有要求的地方就自行处理。看完后删除该说明)

目录

[1 引言 3](#_Toc171285301)

[1.1 背景 3](#_Toc171285302)

[1.2 问题重述 3](#_Toc171285303)

[1.3 工作概述 3](#_Toc171285304)

[1.3.1 问题一分析 3](#_Toc171285305)

[1.3.2 问题二分析 3](#_Toc171285306)

[1.3.3 问题三分析 3](#_Toc171285307)

[1.3.4 问题四分析 3](#_Toc171285308)

[2 一般假设和变量描述 3](#_Toc171285309)

[2.1 假设 3](#_Toc171285310)

[2.2 变量描述 4](#_Toc171285311)

[3 模型的建立与解决方案 4](#_Toc171285312)

[3.1 问题一 4](#_Toc171285313)

[3.1.1 数据预处理 4](#_Toc171285314)

[3.1.2 检验是否满足正态分布 4](#_Toc171285315)

[3.1.3 Spearman相关系数 5](#_Toc171285316)

[3.1.4 Kendall's Tau相关系数 6](#_Toc171285317)

[3.1.5 地理探测器 7](#_Toc171285318)

[3.1.6 分析原因并预防 8](#_Toc171285319)

[3.2 问题二 9](#_Toc171285320)

[3.3 问题三 9](#_Toc171285321)

[3.4 问题四 9](#_Toc171285322)

[4 模型的评估 9](#_Toc171285323)

[4.1 优点 9](#_Toc171285324)

[4.2 缺点 9](#_Toc171285325)

正文内容(原则上不能超过20页)

# 引言

## 背景

洪水通常由暴雨、快速融雪、风暴潮等自然事件引发，导致河流、湖泊的水位急剧上升，是一种常见的自然灾害。洪水也被称为大水，指水体水位超标并威胁相关地区安全，有时甚至会导致灾害发生。随着人口增长，以及为了拓展农田和城市用地，人们对自然环境的破坏（如过度耕作、围湖造田、非法砍伐森林等）改变了地表条件和流水路线，加剧了洪水的严重性。在降水量大的年份，洪水灾害的发生和严重程度往往与人类活动密切相关，其中长期的森林破坏尤为关键。全球气候变化和城市化加速也增加了极端水文事件的频率，使得雨季洪水灾害更加频繁，严重威胁到人民的生命和财产安全[1]。

## 问题重述

1. 首先加载和初步处理train.csv文件，包括数据清洗和处理缺失值。使用皮尔逊相关系数等统计方法来分析20个指标与洪水发生概率之间的相关性，并可视化。根据指标的相关性强度，提出预防洪水的措施和建议。
2. 通过聚类分析，根据洪水发生的概率将数据分为高、中、低三个风险类别。选择合适的指标，并使用决策树、随机森林等方法分析指标重要性并计算权重，从而构建一个洪水风险预警评估模型。最后进行模型的灵敏度分析，评估不同指标权重变化对模型预测结果的影响。
3. 基于问题1的分析结果，选择合适的机器学习模型进行训练。考虑使用特征选择技术来优化模型，尝试只用5个关键指标来调整模型。使用交叉验证等技术验证模型的准确性。
4. 使用问题3中构建的模型，对test.csv中的数据进行洪水概率预测，并将预测结果填充到submit.csv文件中。绘制洪水概率的直方图和折线图，分析结果的分布特征是否正态分布。

## 工作概述

### 问题一分析

本问题主要目的是分析并可视化20个指标中哪些指标与洪水发生概率有密切的相关性。我们的方法首先使用箱形图对原数据进行预处理以排除异常值，接着分析预处理后的数据是否满足正态分布，针对数据的特点选择合适的相关性分析方法。针对数据特点，我们排除了Pearson相关系数分析、Cochran's Q 检验和Kappa一致性检验后尝试使用Spearman相关系数分析、Kendall's Tau相关系数和地理探测器的因子探测器。结合洪水概率洪水概率与其他因素的空间分异性，我们最终使用地理探测器对预处理后的数据进行相关性分析，得到相关性较高和相关性较低的变量，从而针对洪水进行提前预防。

### 问题二分析

对数据进行分割并建立随机森林回归模型，评估其均方误差并分析特征的重要性，从而得出相应的权重。通过这一分析，我们旨在确定

### 问题三分析

### 问题四分析

# 一般假设和变量描述

## 假设

## 变量描述

# 模型的建立与解决方案

## 问题一

### 数据预处理

为了从超过100万条的洪水数据集中识别和处理异常值，我们使用了箱形图。箱形图是一种用于显示数据分布的图表，通过第一四分位数（Q1）、中位数和第三四分位数（Q3）来展示数据的集中趋势和分散程度。箱形图的“箱子”展示了中间50%的数据分布，而从箱子外延伸的“须”则表示数据的极值范围。通常，箱形图外的点被认为是异常值，这些点位于距离箱子1.5倍四分位距（IQR, Q3 - Q1）之外。绘制的箱形图如图 1所示

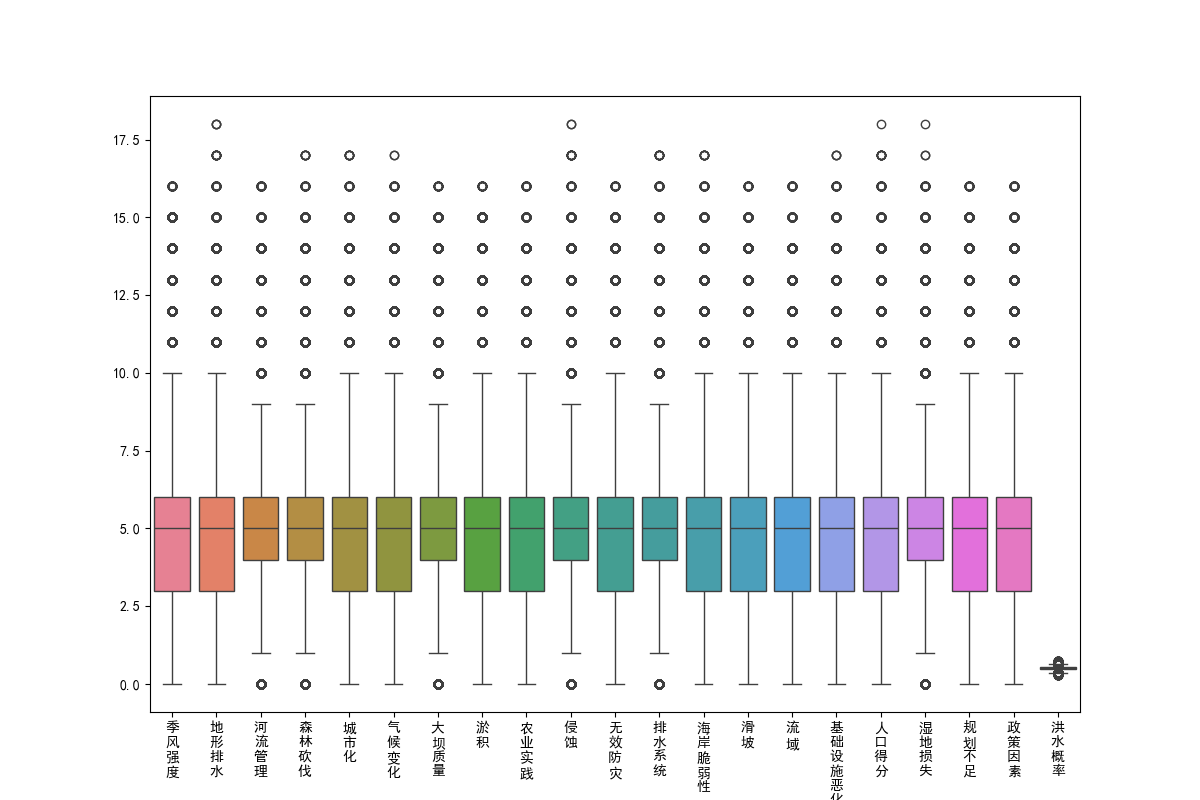


图 1训练集箱形图

根据箱形图识别的异常值，我们计算了四分位距，并依此定义了异常值的范围。数据点若低于Q1减去1.5倍的IQR或高于Q3加上1.5倍的IQR，则被视为异常值并从数据集中剔除。这一步骤是确保后续分析不受极端值影响的关键。

处理完异常值后，对数据集中的数值型特征进行了归一化处理，使用的是最小-最大缩放（MinMaxScaler），将所有特征缩放至[0,1]区间，确保了所有特征在模型训练过程中具有相同的尺度，从而防止某些范围较大的特征在模型中占主导地位，有利于后续的模型建立。

### 检验是否满足正态分布

为了进一步分析数据集中各指标与洪水概率的关系，我们对经过预处理后的数据正态分布的检验。正态分布的假设是许多统计测试的前提条件，包括皮尔逊相关性分析。为了验证数据集中各变量是否符合正态分布，我们应用了Anderson-Darling (AD) 测试，这是一种更为敏感的正态性检验方法，适用于样本大小较大的数据集。与Shapiro-Wilk测试相比，AD测试在样本数量较大时能提供更稳定的结果。我们对数据集中的每个变量执行了AD测试。测试结果包括一个统计量和几个与不同显著性水平相关的临界值。对于每个变量，我们比较了测试统计量与不同显著性水平（如15.0%，10.0%，5.0%，2.5%，1.0%）下的临界值。结果显示，所有变量的测试统计量均明显高于对应的1.0%临界值，表明数据不符合正态分布的假设，我们可以在99%的置信水平下拒绝该数据符合正态分布的假设。

通过前述的Anderson-Darling测试，我们得到train.csv中的数据不符合正态分布，因此不能使用Pearson相关系数分析，此外，由于数据不是二分类数据，我们无法使用Cochran's Q 检验评估二分类数据的组间差异。同样，数据也不是评级数据，无法使用Kappa一致性检验。鉴于以上限制，我们使用非参数的统计方法来分析数据。非参数方法不依赖于数据的分布形式，因此适合处理本研究中的数据类型和分布。Spearman相关系数和Kendall's Tau相关系数均适用于连续或有序的数据，并且能够有效衡量变量之间的单调关系，即使在数据不符合正态分布的情况下也能提供可靠的相关性估计。考虑到洪水概率与多个环境和社会经济因素的复杂关系，我们还选择应用地理探测器，揭示变量的空间分异性对洪水概率的影响。

### Spearman相关系数

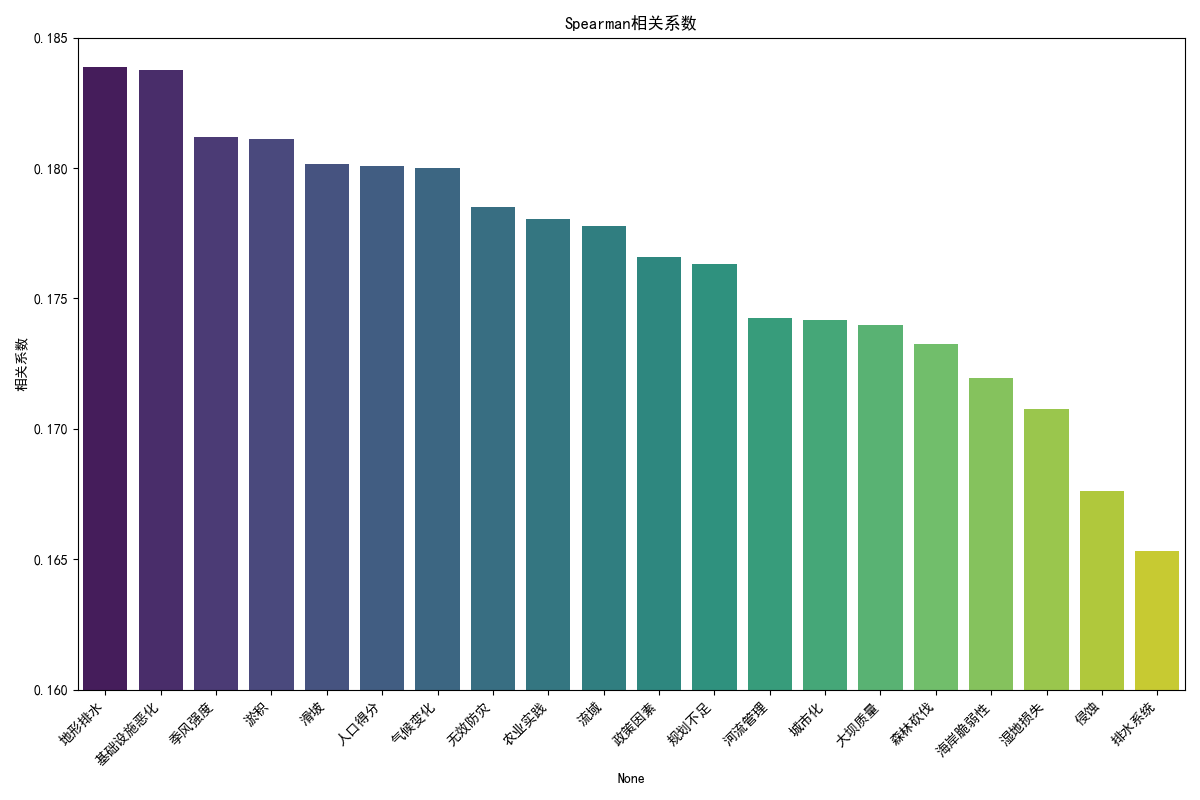
我们使用了Spearman相关系数来评估各变量与洪水概率之间的关系。Spearman相关系数是一种非参数的统计测试，用于衡量两个变量之间的单调关联性。

对于数据集中的每一变量，首先将其与洪水概率变量的数据进行秩次排序，之后计算秩次之间的差异。Spearman相关系数通过以下公式计算：

其中， 是两变量排序位次之差， 是数据点的数量。得到的系数值范围从-1到1，其中1表示完全正相关，-1表示完全负相关，0表示无相关。

将计算得到的相关系数按照与洪水概率的相关强度进行排序，得到一个相关性从高到低的列表。这不仅揭示了各因素对洪水概率影响的强弱，而且帮助识别出对洪水预测模型可能具有重要影响的关键变量。

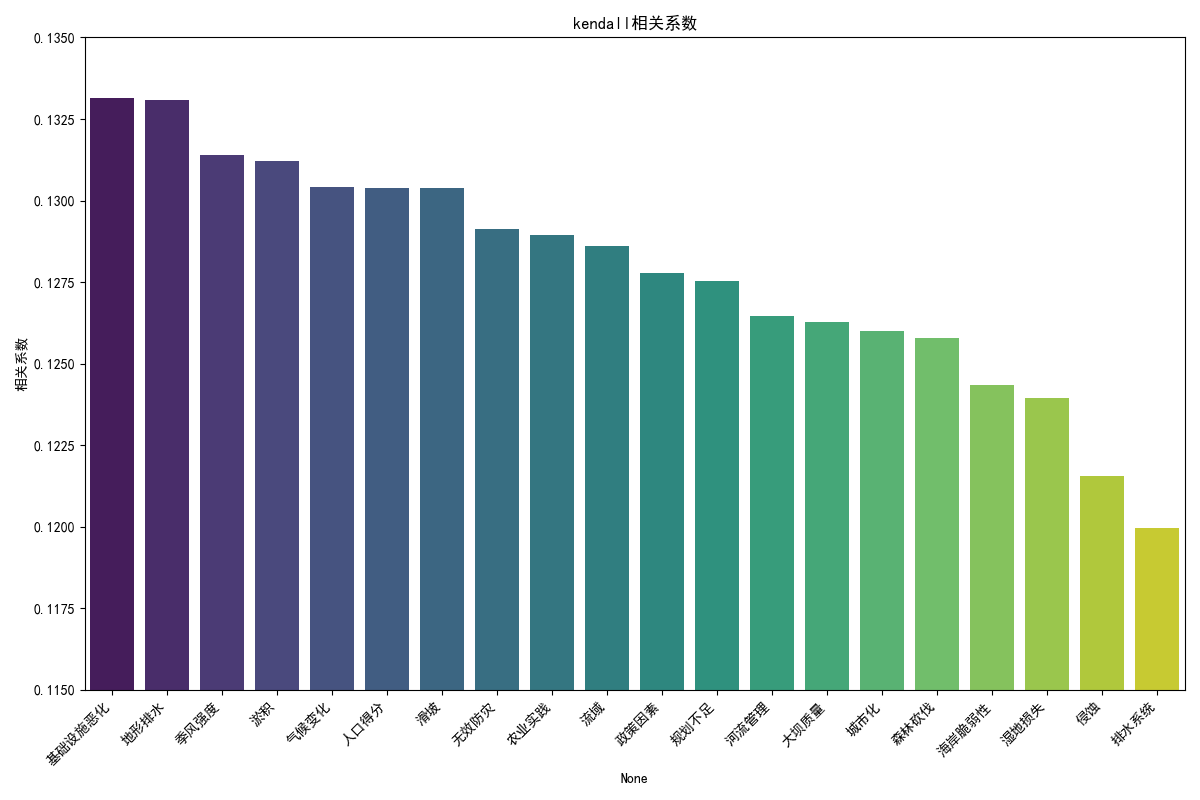
为了直观展示这些相关性，将各变量的相关系数通过柱状图进行展示。图中每个柱子代表一个变量与洪水概率的相关系数，柱状图的高低和颜色变化直观地反映了不同变量的相关强度。



### Kendall's Tau相关系数

为了进一步探索变量间的相关性，我们选择了Kendall's Tau相关系数作对比。Kendall's Tau相关系数是一种用于度量两个变量间有序关联性的非参数统计方法，适合非正态分布数据的分析。具体计算公式为：

其中一致对（concordant pairs）是指在两个变量中，任意两个数据点的排序相同，不一致对（discordant pairs）则排序相反。 是观察对的总数。这个 代表在 个观察值中任取两个数据形成的所有可能对的数量。得到的Kendall's Tau系数提供了一个介于-1到1之间的值，其中1表示完全正相关，-1表示完全负相关，0则表示无相关。通过对所有变量与洪水概率的Kendall's Tau系数进行排序，我们获得了一个反映相关性强度的清晰视图。



### 地理探测器

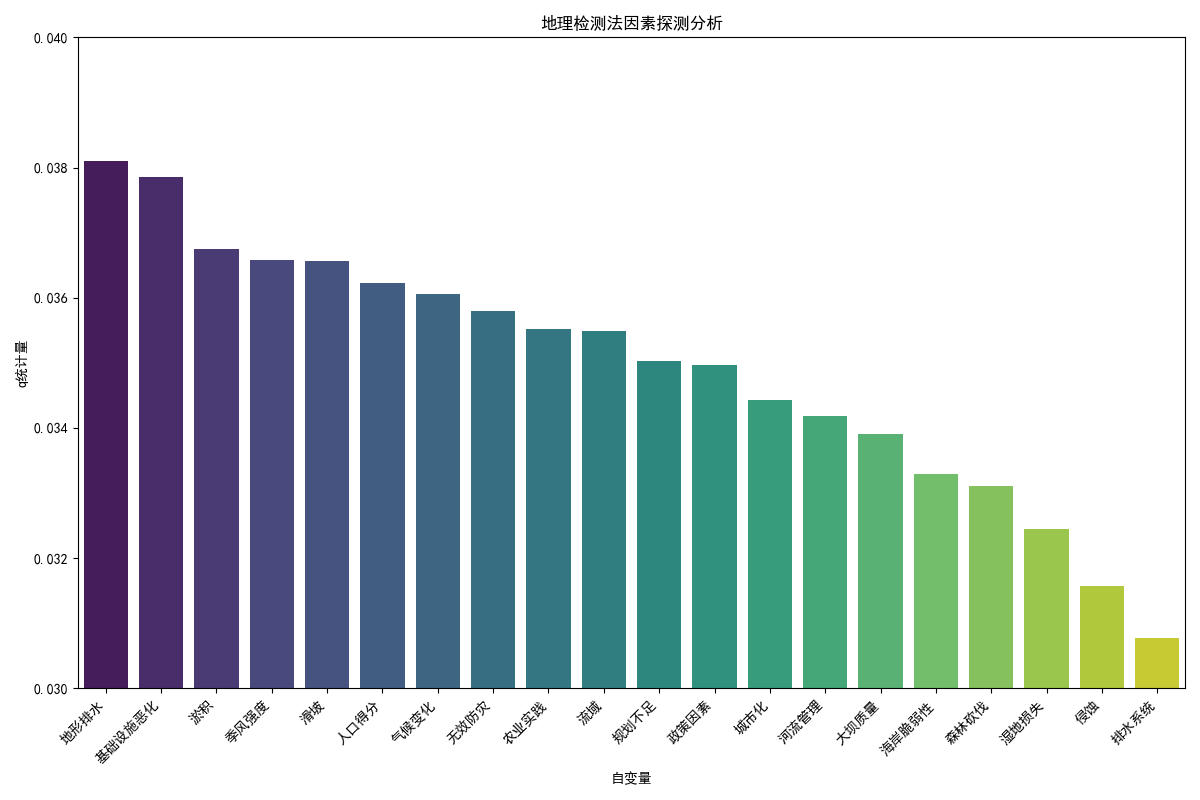
地理探测器是一种用于分析环境因素与事件发生概率之间的空间关联和相互作用,以及揭示不同因素对事件发生的贡献程度的统计学方法。它由中国科学院地理科学与资源研究所的王劲峰研究员提出，并已被广泛应用于社会环境因素和自然环境因素的影响机理研究。地理探测器模型的核心思想是，如果某个自变量对某个因变量有重要影响，那么自变量和因变量的空间分布应该具有相似性。通过计算和比较各单因子的q值，可以判断它们对空间分异性的解释力，q值越大表示解释力越强。表达式为:

其中 为变量 或因子 的分层 (Strata),即分类或分区; 和 分别为层 和全区的单元数; 和 分别是层 和全区的 值的方差。SSW 和 SST 分别为层内方差之和 (Within Sum of Squares) 和全区总方差 (Total Sum of Squares)。 的值域为 ，值越大说明 的空间分异性越明显; 如果分层是由自变量 生成的,则 值越大表示自变量 对属性 的解释力越强,反之则越弱。极端情况下, 值为 1 表明因子 完全控制了 的空间分布, 值为 0 则表明因子 与 没有任何关系, 值表示 解释了 的 。

值的一个简单变换满足非中心 分布

式中: 为非中心参数; 为层 的均值。

通过计算得到的q值，我们对各自变量的贡献程度进行了排序，并通过柱状图形象地展示了这些自变量的影响力。柱状图显示了自变量对洪水概率的解释程度，其中每个柱子的高度代表对应变量的q统计量。



由于地理探测器q 值具有明确的物理含义，没有线性假设，客观地探测出自变量解释了100×q%的因变量。通过空间异质性来探测因变量与自变量之间空间分布格局的一致性，据此度量自变量对因变量的解释度，即q 值。因其不基于线性假设，提供了更为强有力的证据支持它们与洪水发生可能存在的因果关系。

对此，我们选择q值在0.036以上的指标认定为该指标与洪水的发生有着密切的关联，即地形排水、基础设施恶化、淤积、季风强度、滑坡、人口得分和气候变化，q值在0.036以下的指标认定为该指标与洪水发生相关性不大。

### 分析原因并预防

针对地理探测器得出的密切相关因素，结合自然地理原理，分析上述指标相关性大的可能的原因是①地形排水不良可能导致雨水积聚，增加表面径流，导致洪水发生的概率增加。地势低洼或斜坡陡峭，可能阻碍有效的水流排出，造成水体在特定区域的积累。②老化或破损的基础设施降低了地区对洪水的防御能力。此外，缺乏适当的维护可能导致防洪系统功能衰退，无法在雨季有效控制水流。③河流和水库的淤积降低了它们的储水能力和调节流量的能力。在强降水时，这些水体无法容纳额外的水流，从而增加了洪水溢出的风险。④强烈的季风带来的高降水量直接增加了洪水的发生几率。季风导致的持续降雨在短时间内显著增加地表水和河流水位。⑤滑坡可以暂时阻塞河流，形成自然堤坝，当这些堤坝在水压作用下突然决堤时，将导致大量积水瞬间下泄，形成洪水。⑥高人口密度地区通常伴随高度城市化，地面硬化和绿地减少，导致自然地面吸水能力降低，增加洪水风险。⑦气候变化导致的极端天气事件频发，如极端降雨和风暴潮，增加了洪水事件的发生频率和强度。

对此，我们作出如下建议：优化城市和乡村的排水系统设计，增设雨水收集和排放设施。定期检查和维护防洪堤和排水系统，更新老化的基础设施。定期清理河床，减少河流淤积，确保河道流通，增加水体的调蓄能力。建立健全的预警系统，季风来临前进行风险评估和应急预案的准备。在滑坡易发区域进行地质评估，采取工程措施稳定斜坡，监控重点区域的地质活动。在城市规划中考虑洪水风险，限制在洪水易发区的建设，增加绿地和透水表面。提高对气候变化影响的适应能力，投资于可持续的水资源管理和保护方案。

## 问题二

## 问题三

## 问题四

# 模型的评估

## 优点

## 缺点

**参考文献**

1. 张建云,宋晓猛,王国庆等.变化环境下城市水文学的发展与挑战：I.城市水文效应[J].水科学进展,2014,25(4):594-605.
2. ZHANG Jianyun,SONG Xiaomeng,WANG Guoqing, et al.Development and challenges of urban hydrology in a changing environment:I: Hydrological response to urbanization[J]. Advances in Water Science, 2014,25(4):594-605.
3. AINUDDIN S, ROUTRAY J K, AINUDDIN S. People's risk perception in earthquake prone Quetta city of Baluchistan[J]. International Journal of Disaster Risk Reduction, 2014, 7: 165-175.
4. TABARI H. Climate change impact on flood and extreme precipitation increases with water availability[J]. Scientific Reports, 2020, 10(1): 13768.
5. **王劲峰,徐成东.地理探测器:原理与展望[J].地理学报,2017,72(01):116-134.**

参考文献的编号，如[1][3]等；引用书籍还必须指出页码。参考文献按正文中的引用次序列出，其中：**书籍的表述方式为**

[编号] 作者，书名，出版地：出版社，出版年。

**参考文献中期刊杂志论文的表述方式为**

[编号] 作者，论文名，杂志名，卷期号：起止页码，出版年。

**参考文献中网上资源的表述方式为**

[编号] 作者，资源标题，网址，访问时间（年月日）。

**附录（另起一页）**