

基于K-Means聚类等综合模型对洪水灾害的数据分析与预测

邵建鑫

广州工商学院通识教育学院, 广东 广州

收稿日期: 2025年4月28日; 录用日期: 2025年5月21日; 发布日期: 2025年5月30日

摘要

在当今社会, 全球气候变化对人类生活和环境造成了日益严重的影响, 特别是频发的极端天气事件如洪水给社会经济发展带来了巨大挑战。洪水不仅造成了财产损失和人员伤亡, 还对基础设施、农业生产和生态环境造成长期影响。因此, 准确评估洪水风险并采取有效的预防和应对措施至关重要。本文基于K-means聚类和支撑向量机(SVM)等方法, 分析了洪水发生的多因素影响, 并建立了洪水风险评估模型。首先, 我们通过灰色关联分析和数据挖掘技术, 识别了影响洪水发生的关键因素, 并对这些因素进行了量化和分析。其次, 我们应用K-means聚类方法对不同风险级别的洪水进行了分类和评估。最后, 通过支撑向量机(SVM)模型, 我们预测了洪水发生概率, 并评估了模型的预测准确性和稳定性。

关键词

K-Means聚类, 洪水灾害, 灰色关联分析

Data Analysis and Prediction of Flood Disasters Based on Comprehensive Models Such as K-Means Clustering

Jianxin Shao

General Education College, Guangzhou College of Technology and Business, Guangzhou Guangdong

Received: Apr. 28th, 2025; accepted: May 21st, 2025; published: May 30th, 2025

Abstract

In today's society, global climate change has exerted increasingly severe impacts on human life and the environment, particularly with the frequent occurrence of extreme weather events such as floods,

文章引用: 邵建鑫. 基于 K-Means 聚类等综合模型对洪水灾害的数据分析与预测[J]. 应用数学进展, 2025, 14(5): 580-588. DOI: 10.12677/aam.2025.145284

which pose significant challenges to socio-economic development. Floods not only cause property damage and casualties, but also have long-term effects on infrastructure, agricultural production, and the ecological environment. Therefore, accurately assessing flood risks and implementing effective prevention and response measures are of paramount importance. This paper analyzes the multi-factor influences on flood occurrence and establishes a flood risk assessment model based on methods such as K-means clustering and Support Vector Machine (SVM). First, we identify key factors affecting flood occurrence through grey relational analysis and data mining techniques, quantifying and analyzing these factors. Second, we apply the K-means clustering method to classify and evaluate floods of different risk levels. Finally, using the Support Vector Machine (SVM) model, we predict the probability of flood occurrence and assess the model's predictive accuracy and stability.

Keywords

K-Means Clustering, Flood Disaster, Grey Correlation Analysis

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 研究背景

洪水灾害的数据分析与预测[1]

洪水是暴雨、急剧融冰化雪、风暴潮等自然因素引起的江河湖泊水量迅速增加，或者水位迅猛上涨的一种自然现象，是自然灾害。洪水又称大水，是河流、海洋、湖泊等水体上涨超过一定水位，威胁有关地区的安全，甚至造成灾害的水流。洪水一词，在中国出自先秦《尚书·尧典》。从那时起，四千多年中有过很多次水灾记载，欧洲最早的洪水记载也远在公元前 1450 年。在西亚的底格里斯 - 幼发拉底河以及非洲的尼罗河关于洪水的记载，则可追溯到公元前 40 世纪。2023 年 6 月 24 日 8 时至 25 日 8 时，中国 15 条河流发生超警洪水。2023 年，全球洪水等造成了数十亿美元的经济损失。

洪水的频率和严重程度与人口增长趋势相当一致。迅猛的人口增长、扩大耕地、围湖造田、乱砍滥伐等人为破坏不断地改变着地表状态，改变了汇流条件，加剧了洪灾程度。在降水多的年份，洪水是否造成灾害，以及洪水灾害的大小，也离不开人为因素，长期以来人为的森林破坏是其重要原因。长江上游乱砍滥伐的恶果是惊人的水土流失。现已达 35 万平方千米，每年土壤浸蚀量达 25 亿吨。河流、湖泊、水库淤积的泥沙量达 20 亿吨。仅四川一省一年流入长江各支流的泥沙，如叠成宽高各 1 米的堤，可以围绕地球赤道 16 圈。我国第一大淡水湖洞庭湖每年沉积的泥沙达 1 亿多吨，有专家惊呼：“这样下去，要不了 50 年，洞庭湖将从地球上消失！”长江之险，险在荆江，由于泥沙俱下，如今荆江段河床比江外地面高出十多米，成了除黄河之外名副其实的地上河。对森林的肆意砍伐不仅危害自己，而且祸及子孙后代，世界上许多地方，如美索不达米亚、小亚细亚、阿尔卑斯山南坡等由于过度砍伐森林，最后都变成了不毛之地[2]。

- 1) 将洪水发生的概率聚类成不同类别，分析具有高、中、低风险的洪水事件的指标特征。
- 2) 选取合适的指标，计算不同指标的权重，建立发生洪水不同风险的预警评价模型。
- 3) 最后进行模型的灵敏度分析。
- 4) 政策制定：制定防洪减灾相关政策，加强洪水预警和应急响应系统。

需要我们对洪水概率数据进行聚类分析，将洪水事件分为高、中、低风险类别。分析不同类别洪水事件的指标特征，计算各指标的权重，建立预警评价模型，并进行模型的灵敏度分析。

- 1) 聚类分析：使用层次聚类算法，根据洪水发生概率将洪水事件分为高、中、低风险类别。

- 2) 特征分析：分析不同风险类别的指标特征，找出高风险类别中显著的指标特征。
- 3) 权重计算：使用因子分析，计算各指标在不同风险类别中的权重。
- 4) 预警评价模型：建立加权评分模型，对每个洪水事件进行风险评估和预警。
- 5) 灵敏度分析：通过调整模型参数，评估模型对不同参数变化的敏感性，确保模型的稳定性和可靠性。

2. 数学模型

基于 K-means 聚类的洪水风险聚类与基于灰色关联分析的预警评价模型[3]，模型符号说明见表 1。

Table 1. Explanation of terminology symbols

表 1. 术语符号说明

术语/符号	定义/说明
x_1	季风强度：反映季风对洪水发生的影响。
x_2	地形排水：反映地形和排水条件对洪水的影响。
x_3	河流管理：反映河流管理措施对洪水的影响。
x_4	森林砍伐：反映森林砍伐对洪水的影响。
x_5	城市化：反映城市化水平对洪水的影响。
x_6	气候变化：反映气候变化对洪水的影响。
x_7	大坝质量：反映大坝质量对洪水的影响。
x_8	淤积：反映淤积程度对洪水的影响。
x_9	农业实践：反映农业活动对洪水的影响。
x_{10}	侵蚀：反映土壤侵蚀对洪水的影响。
x_{11}	无效防灾：反映防灾措施的有效性对洪水的影响。
x_{12}	排水系统：反映排水系统对洪水的影响。
x_{13}	海岸脆弱性：反映海岸脆弱性对洪水的影响。
x_{14}	滑坡：反映滑坡对洪水的影响。
x_{15}	流域：反映流域特征对洪水的影响。
x_{16}	基础设施恶化：反映基础设施状况对洪水的影响。
x_{17}	人口得分：反映人口密度对洪水的影响。
x_{18}	湿地损失：反映湿地损失对洪水的影响。
x_{19}	规划不足：反映规划不足对洪水的影响。
x_{20}	政策因素：反映政策对洪水的影响。
$y_i, i = 1, 2, \dots, 20$	每个指标 x_i 对应得到分样本下的洪水概率发生值
x	总样本下的指标值
y	总样本下的洪水概率发生值
$\bar{x}$	总样本下各指标的均值。
$\bar{y}$	总样本下各指标对应洪水发生的均值。
r_{xy}	总样本下各指标与洪水发生概率下的皮尔逊相关系数
$\beta_0, \beta_1, \dots, \beta_{20}$	各个指标对应的线性回归系数
ε	线性回归模型的误差项
$P(Y = 1)$	洪水是否发生的概率

聚类分析模型采用 K-means 聚类算法对洪水概率进行聚类, 以确定洪水风险的高、中、低类别。K-means 聚类是一种基于迭代优化的非监督学习方法, 其数学原理可以通过以下步骤描述:

1) 初始步骤: 选择 k 个初始聚类中心(即质心), 通常随机选取。

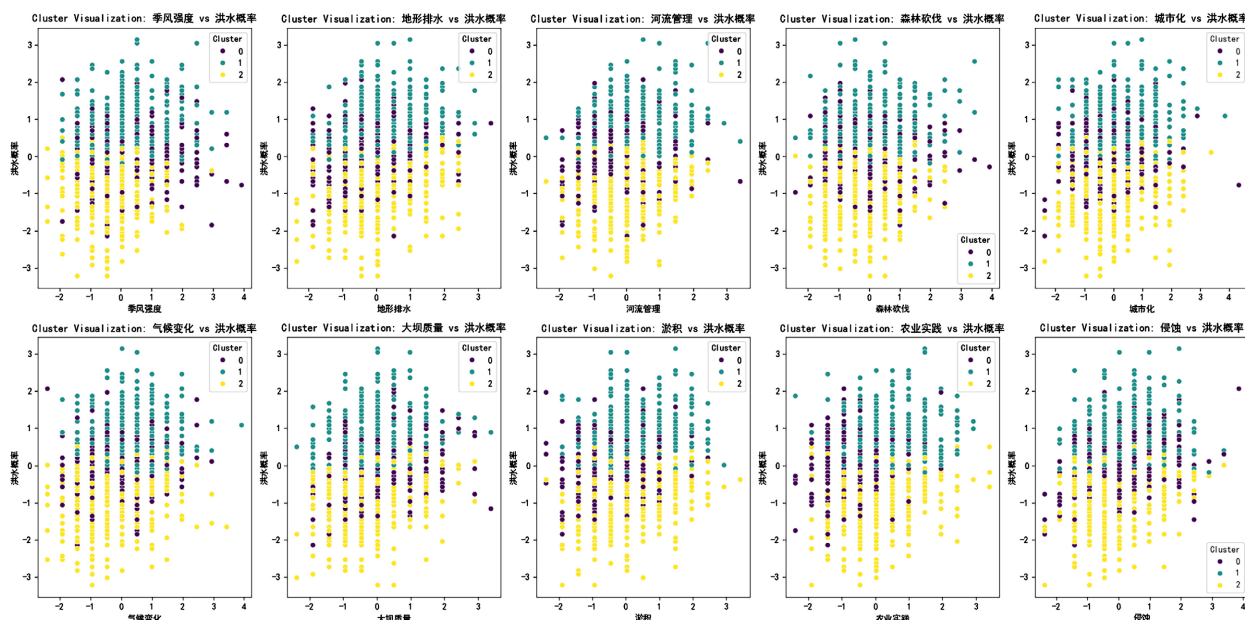


Figure 1. Clustering results of flood risk between the top 10 indicators and flood occurrence probability

图 1. 前 10 个指标与洪水发生概率之间的洪水风险聚类结果

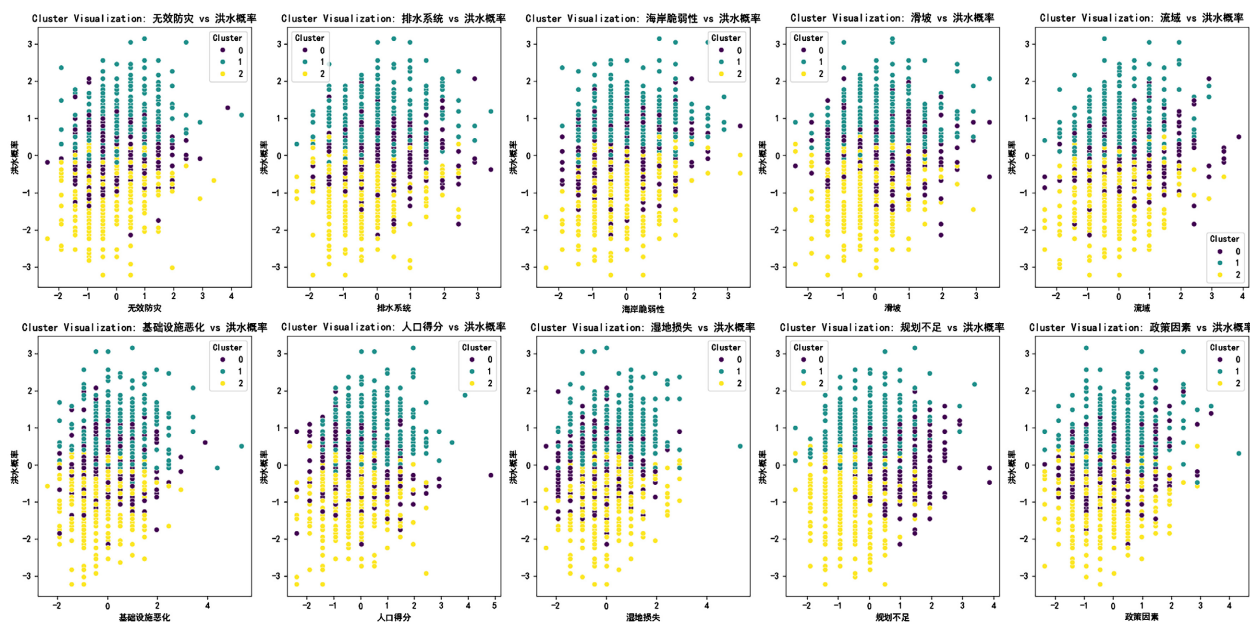


Figure 2. Clustering results of flood risk between the latter 10 indicators and flood occurrence probability

图 2. 后 10 个指标与洪水发生概率之间的洪水风险聚类结果

2) 分配步骤: 对于数据集中的每个数据点 x_i , 计算其与每个质心 μ_j 的距离, 将 x_i 分配到距离最近的质心所在的簇。距离通常使用欧氏距离衡量, 即:

$$d(x_i, \mu_j) = \sqrt{\sum_{l=1}^m (x_{il} - \mu_{jl})^2}$$

其中, m 是数据的维度, x_{il} 和 μ_{jl} 分别表示数据点 x_i 和质心 μ_j 在第 l 个维度的值。

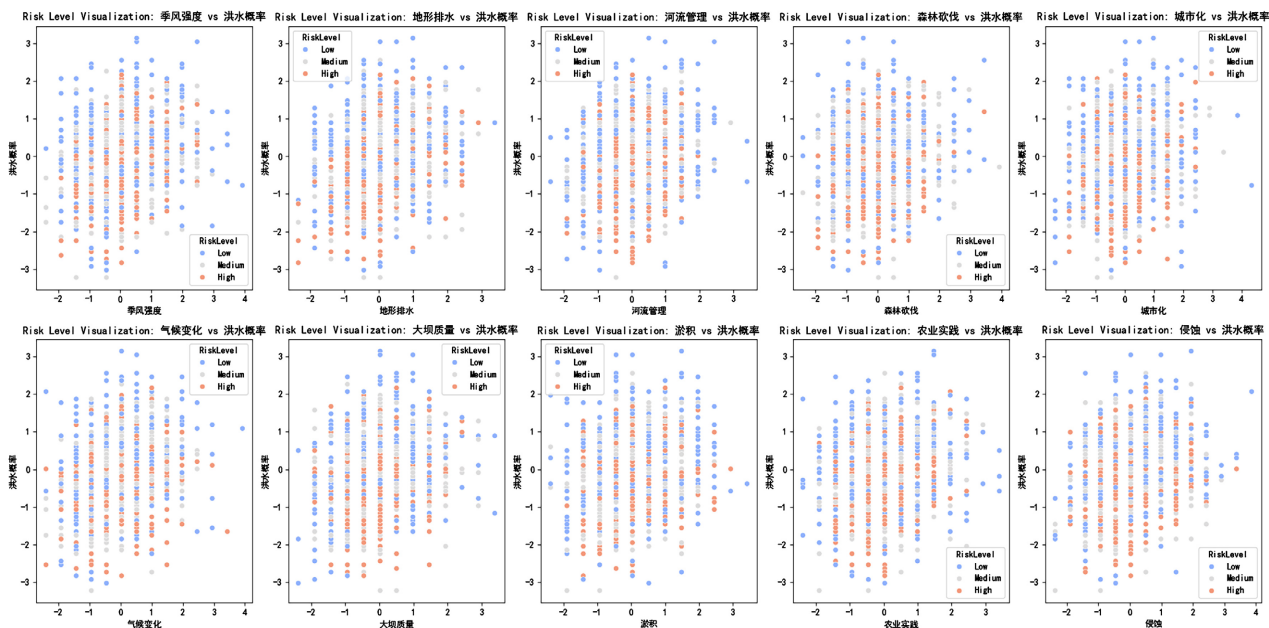


Figure 3. Evaluation results of flood risk early warning between the top 10 indicators and flood occurrence probability
图 3. 前 10 个指标与洪水发生概率之间的洪水风险预警评价结果

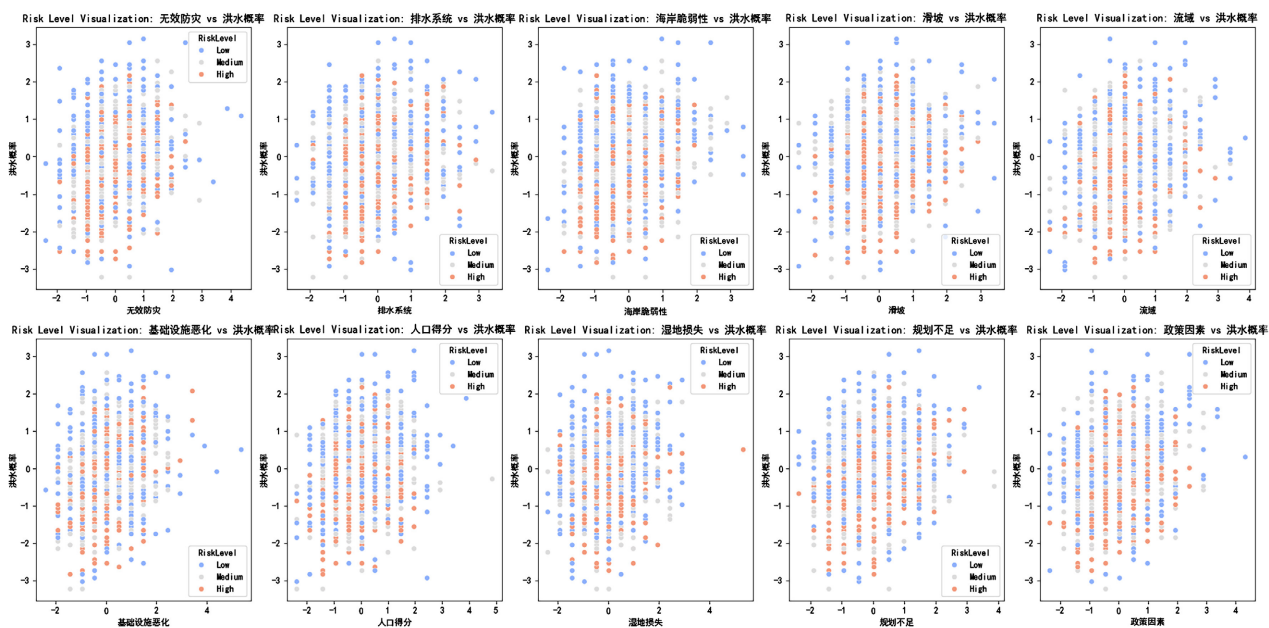


Figure 4. Evaluation results of flood risk early warning between the latter 10 indicators and flood occurrence probability
图 4. 后 10 个指标与洪水发生概率之间的洪水风险预警评价结果

3) 更新步骤: 计算每个簇的新质心。新质心是簇中所有数据点的均值, 即:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

其中, C_j 表示第 j 个簇的集合, $|C_j|$ 是簇中数据点的数量。

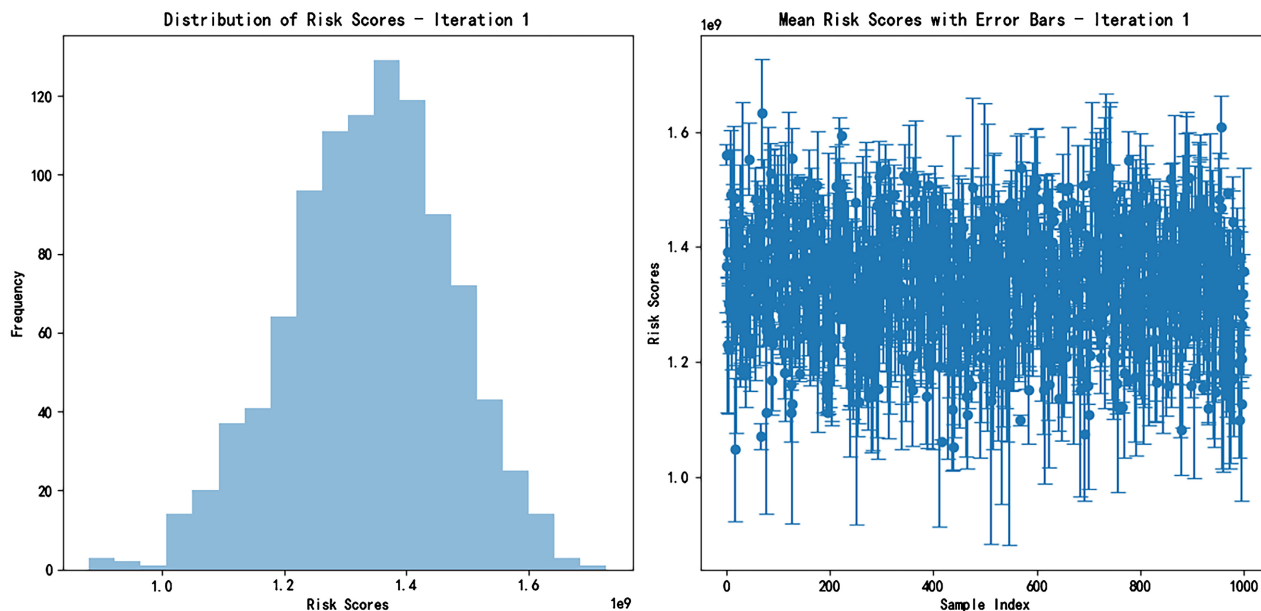


Figure 5. Output score distribution and error bar chart of the flood risk early warning evaluation model between indicators and flood occurrence probability

图 5. 指标与洪水发生概率之间的洪水风险预警评价模型输出评分分布图和误差条形图

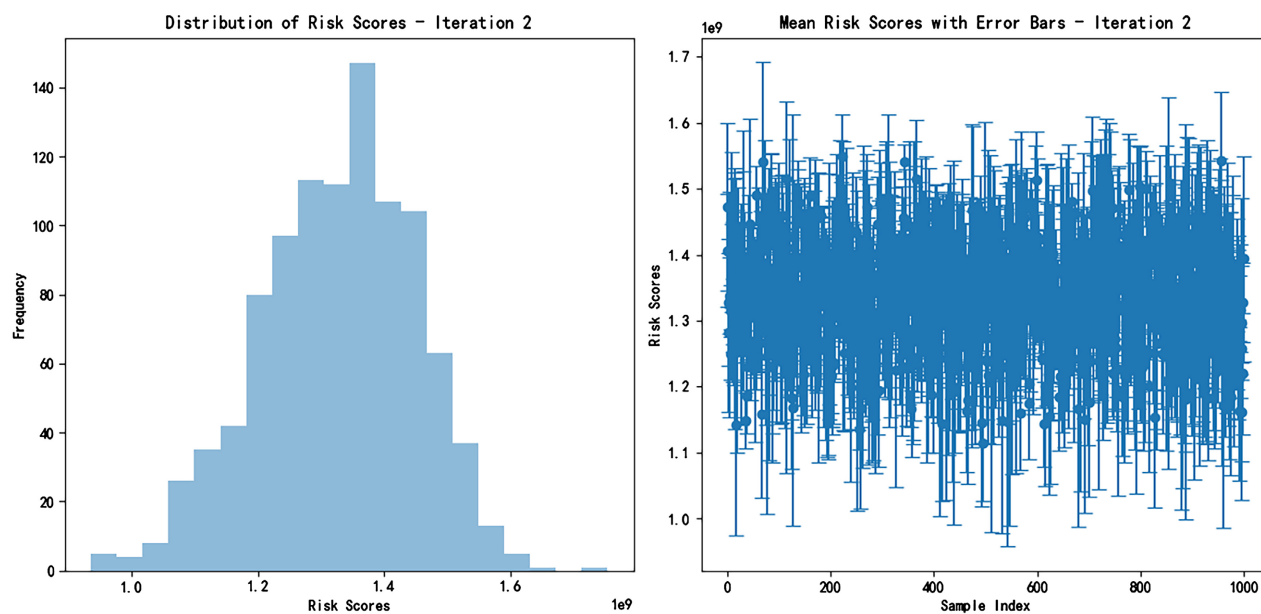


Figure 6. Output score distribution and error bar chart of the flood risk early warning evaluation model between indicators and flood occurrence probability.

图 6. 指标与洪水发生概率之间的洪水风险预警评价模型输出评分分布图和误差条形图

4) 迭代步骤: 重复分配步骤和更新步骤, 直到质心位置不再变化或变化小于设定的阈值。K-means

聚类的目标是 최소화所有数据点到其所属簇质心的距离平方和，即：

$$J = \sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

其中， J 是损失函数， k 是聚类的数量， C_j 是第 j 个簇， μ_j 是第 j 个簇的质心。通过 Python 绘图，我们得到如图 1、图 2 所示的洪水风险聚类结果。

5) 基于灰色关联分析的洪水预警评价模型：灰色关联分析是一种多变量分析方法，适用于处理数据不完备或信息不确切的情况。在洪水预警评价模型中，符号说明见表 1，可以通过以下步骤应用灰色关联分析[4]：

① 数据准备：选择关键因素并对数据进行归一化处理，以确保不同因素处于相同的量纲和尺度。

② 灰色关联度计算：对于第 j 个因素，计算其与洪水发生概率序列 $\{x_{1j}, x_{2j}, \dots, x_{nj}\}$ 的关联度。灰色关联度 ξ_j 可通过以下公式计算：

$$\xi_j = \frac{\sum_{i=1}^{n-1} \rho(|x_{ij+1} - x_{ij}|)}{\sum_{i=1}^{n-1} |x_{ij+1} - x_{ij}| + \varepsilon}$$

其中， $\rho(\cdot)$ 是关联函数， ε 是平滑系数，通常取一个很小的正数，以避免分母为零的情况。

③ 确定权重：将计算得到的灰色关联度 ξ_j 标准化为权重 ω_j ，使得各因素的权重之和为 1。

即：

$$\omega_j = \frac{\xi_j}{\sum_{j=1}^m \xi_j}$$

这里， m 是因素的总数。

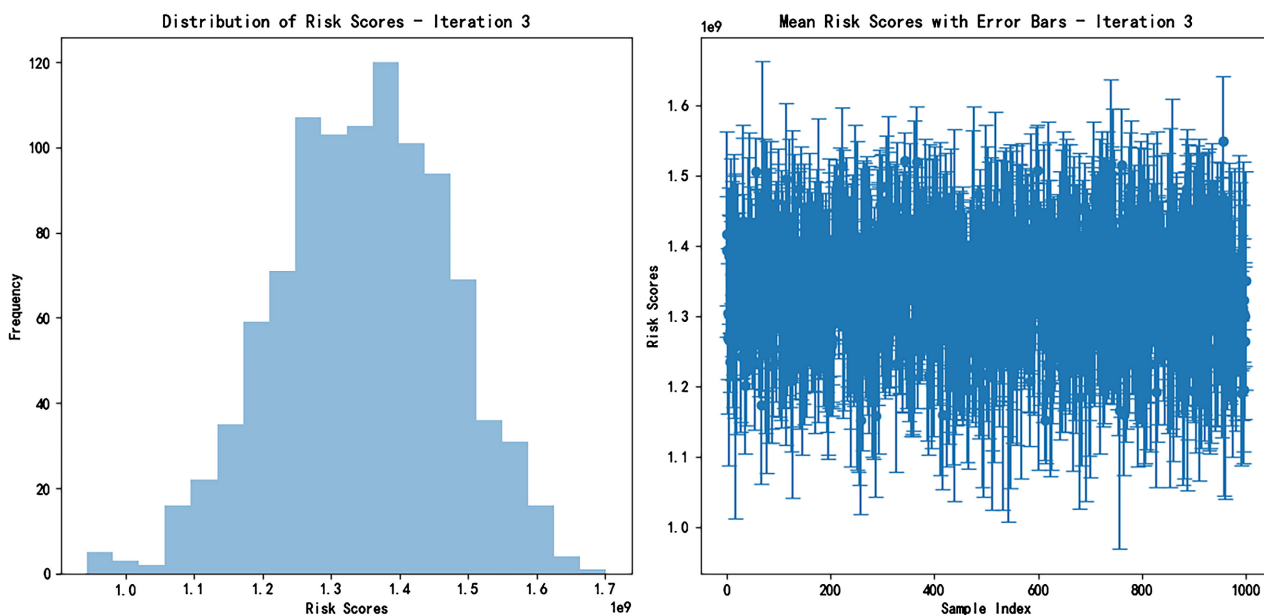


Figure 7. Output score distribution and error bar chart of the flood risk early warning evaluation model between indicators and flood occurrence probability after repeated runs

图 7. 重复运行后指标与洪水发生概率之间的洪水风险预警评价模型输出评分分布图和误差条形图

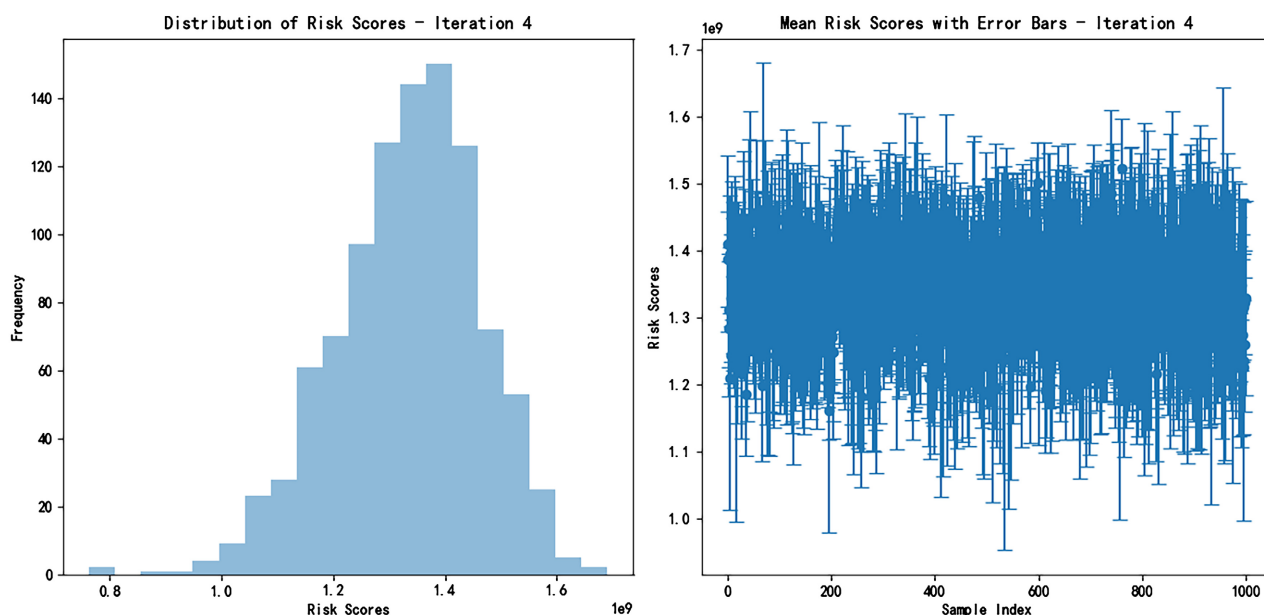


Figure 8. Output score distribution and error bar chart of the flood risk early warning evaluation model between indicators and flood occurrence probability after repeated runs

图 8. 重复运行后指标与洪水发生概率之间的洪水风险预警评价模型输出评分分布图和误差条形图

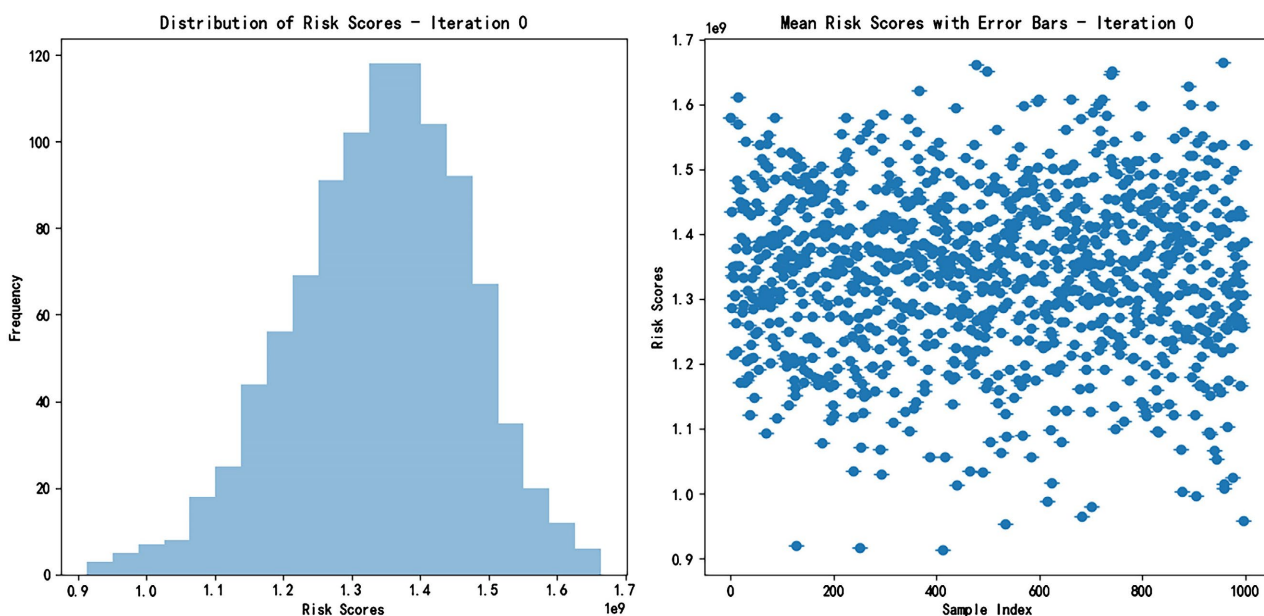


Figure 9. Distribution diagram or error bar chart results of the flood risk early warning evaluation model output scores between indicators and flood occurrence probability after running the model repeatedly for more than 5 times

图 9. 重复运行模型 5 次以上，指标与洪水发生概率之间的洪水风险预警评价模型输出评分的分布图或误差条形图结果

④ 风险等级评估：综合各因素的权重和实际数据，评估洪水的风险等级。高灰色关联度对应高风险，低灰色关联度对应低风险[5]。

通过灰色关联分析，我们可以更准确地评估洪水事件的风险等级，并提供相应的预警建议和管理措施，以降低潜在的洪水影响。

通过 Python 画图我们得到如图 3、图 4 所示的洪水风险预警评价结果。

由图 3、图 4，我们容易看出基础设施恶化、地形排水、排水系统、季风强度、气候变化这些指标是洪水高风险预警评价等级最相关的指标，人口得分、农业实践、海岸脆弱性这些指标是洪水低风险预警评价等级最不相关的指标。

灵敏度分析：为保证模型输出稳定性，我们重复运行模型 5 次以上，比较多次运行的结果，分析模型输出的一致性和稳定性。绘制模型输出评分的分布图或误差条形图，以展示模型在不同运行中输出结果的变化情况。通过 Python 画图，我们得到如图 5~9 所示的洪水风险预警评价模型输出评分的分布图或误差条形图结果(这里只展示最佳的一次结果，更多结果见[6])。

由图 9，容易观察得到我们的模型灵敏度较高，重复运行模型 5 次以上之后，指标与洪水发生概率之间的洪水风险预警评价模型输出评分的分布符合正态分布，误差较小。

3. 研究结论

基于 K-means 聚类和灰色关联分析的基础理论模型[7]，我们建立了具体的洪水风险聚类 and 预警评价结果，如图 1~4 所示，通过这些图的结果，我们发现基础设施恶化、地形排水、排水系统、季风强度、气候变化这些指标是洪水高风险预警评价等级最相关的指标，人口得分、农业实践、海岸脆弱性这些指标是洪水低风险预警评价等级最不相关的指标。

另外，为保证模型输出稳定性，我们重复运行模型 5 次以上，比较多次运行的结果，分析模型输出的一致性和稳定性。绘制模型输出评分的分布图或误差条形图，以展示我们的模型在不同运行中输出结果的变化情况。

得到如图 5~9 所示的洪水风险预警评价模型输出评分的分布图或误差条形图结果。并由此说明我们的模型灵敏度较高，重复运行模型 5 次以上之后，指标与洪水发生概率之间的洪水风险预警评价模型输出评分的分布符合正态分布[8]，误差较小，说明我们的模型拟合度较高、较合理。

参考文献

- [1] 高思如. 震害预测中强震溃坝洪水灾害分析评估系统与坝-坝灾害链研究[D]: [硕士学位论文]. 兰州: 中国地震局兰州地震研究所, 2011.
- [2] 魏国忠, 曾伟. 用灰色系统理论分析预测滹沱河洪水灾害[J]. 河北水利, 2003(4): 36-38.
- [3] 管华. 淮河上游洪水灾害可公度信息系预测[J]. 灾害学, 1996(4): 21-24.
- [4] 张帆, 张永勇, 陈俊旭, 等. 多种机器学习模型对不同洪水类型特征指标模拟效果评估[J]. 地理科学进展, 2022, 41(7): 1239-1250.
- [5] 魏一鸣, 范英, 傅继良, 等. 基于神经网络的洪水灾害预测方法[J]. 中国管理科学, 2000, 8(3): 58-63.
- [6] 巴欢欢, 郭生练, 钟逸轩, 等. 考虑降水预报的三峡入库洪水集合概率预报方法比较[J]. 水科学进展, 2019, 30(2): 186-197.
- [7] 沈婕, 梁忠民, 胡义明, 等. Copula-MCP 洪水概率预报方法研究与应用[J]. 中国农村水利水电, 2020(1): 125-129.
- [8] 王定成, 汪春秀, 朱天一, 等. 基于 SVM 的灾害天气预测方法的研究[J]. 武汉理工大学学报, 2010, 32(24): 121-124.