

林森, 刘蓓蓓, 闫雪, 等. 灾害大数据驱动的县域重大洪涝过程灾害风险评估[J]. 灾害学, 2022, 37(4): 166–172.  
[LIN Sen, LIU Beibei, YAN Xue, et al. Disaster Risk Assessment at County Level of a Heavy Flooding Driven by Disaster Big Data [J]. Journal of Catastrophology, 2022, 37(4): 166–172. doi: 10.3969/j.issn.1000-811X.2022.04.026.]

# 灾害大数据驱动的县域重大洪涝过程灾害风险评估<sup>\*</sup>

林 森, 刘蓓蓓, 闫 雪, 孙 宁, 郭桂祯

(应急管理部国家减灾中心, 北京 100124)

**摘 要:** 基于我国南方地区 625 个重大洪涝过程案例, 收集了县域 23 项指标的近 30 万条数据, 利用 XGBoost 算法建立了 6 个重大洪涝过程灾害风险评估模型, 用于灾害事件发生前对受灾人口风险、紧急转移安置人口风险、农作物受灾面积风险、倒塌和严重损坏房屋风险、直接经济损失风险和灾害综合风险进行评估。通过实际案例对模型进行验证, 灾害风险评估结果准确率整体达到 80% 以上, 表明该模型具有较好的泛化能力, 能够用于实际灾害评估工作中。实验对比发现, 全指标比仅用致灾因子指标可以使评估准确率提升 10% ~ 15%; 另外, 训练样本量提高 1 ~ 2 个数量级也能够使模型评估准确率提升 5% ~ 13%, 这表明灾害大数据的积累对灾害风险评估工作具有重要意义。

**关键词:** 洪涝灾害; 大数据; 风险评估; 评估指标; XGBoost 算法

**中图分类号:** X43; X915.5; P642 **文献标志码:** A **文章编号:** 1000-811X(2022)04-0166-07

doi: 10.3969/j.issn.1000-811X.2022.04.026

根据应急管理部发布数据, 2021 年我国重大洪涝灾害共造成 5 901 万人次受灾, 因灾死亡失踪 590 人, 倒塌房屋 15 万间, 直接经济损失 2 459 亿元。其中, 华北、西北地区洪涝灾害历史罕见, 特别是 7 月份河南遭遇特大暴雨洪涝灾害, 造成近 1 500 万人受灾, 直接经济损失 1 200 亿元<sup>[1]</sup>。随着全球气候变化、极端气候事件频发和社会经济的快速发展, 暴雨多、强度大、损失重已成为洪涝灾害的趋势<sup>[2-3]</sup>。减少洪涝灾害损失, 减轻洪涝灾害风险是我国洪涝灾害风险管理中的必要任务。

风险评估是风险管理的重要基础。根据洪涝灾害风险评估的时间段, 可分为针对区域历史灾害的风险分析和针对实时重大洪涝过程的风险预测两大类。前者可用于风险区划和洪水风险图编制, 帮助政府制定长期规划, 主要方法是通过指标体系对区域风险进行评估, 一般选择致灾因子、承灾体、孕灾环境、防灾减灾能力等灾害系统相关指标<sup>[4-5]</sup>, 经过层次分析法、熵权法等方法确定指标权重, 进行加权求得一个综合风险指数, 在此基础上进行风险等级划分, 此类方法主观性较强<sup>[6-7]</sup>, 结果以定性评价为主, 另外还有模型模拟<sup>[8-9]</sup>、信息扩散<sup>[10-11]</sup>等方法。针对实时重大洪涝过程的灾害风险评估研究相对较少, 其目的是对即将发生的洪涝过程预测承灾体期望损失和灾害风险分布, 为灾害预警和应急响应提供依据。目前, 国际上单灾种风险评估模型以半定量化和

定量化模型居多, 如美国的自然灾害风险评估模型(HAZUS)<sup>[12]</sup>、世界银行的概率风险评估模型(CAPRA)<sup>[13]</sup>等, 这些模型大多基于致灾因子危险性(H)、承灾体脆弱性(V)、承灾体暴露度(E)三要素<sup>[14]</sup>。对洪涝灾害而言, 一般是将不同强度暴雨概率、承灾体脆弱性函数、承灾体分布三者进行耦合, 构建定量化风险评估模型<sup>[15-16]</sup>, 用于评估单次洪涝过程可能造成的人口、房屋、农作物、经济等期望损失<sup>[17]</sup>。基于 H-V-E 框架的洪涝风险评估模型核心是计算不同暴雨概率和开展尺度适宜的脆弱性函数研究<sup>[18-21]</sup>, 该方法在应用中可能存在三个方面的困难: ①尺度太小会导致样本量小, 难以获取较好的回归函数, 影响脆弱性函数表现<sup>[22]</sup>, 而尺度太大会导致脆弱性代表性不足; ②建模复杂, 需要分步骤开展危险性和脆弱性研究, 而且不同强度危险性分级会造成计算结果离散化, 不能形成连续度量; ③难以考虑到灾害系统中其他指标, 如孕灾环境指标、防灾减灾能力指标等<sup>[23]</sup>。

大数据背景下, 数据驱动的机器学习算法应用逐渐增多, 很多算法被用于建立灾害风险评估模型, 其优势是能够利用大样本数据进行建模<sup>[24-25]</sup>, 建模过程淡化了传统灾害风险理论中危险性、脆弱性等概念, 简化了模型搭建步骤, 同时, 还可以考虑更多的相关指标。例如, 刘芳利

<sup>\*</sup> 收稿日期: 2022-05-15 修回日期: 2022-08-31

基金项目: 国家重点研发计划资助(2018YFC1508806)

第一作者简介: 林森(1987-), 男, 汉族, 山东青岛人, 博士, 助理研究员, 主要从事自然灾害评估、机器学习算法应用等方面研究。

E-mail: linsen\_ndrcc@126.com

用降雨量、孕灾环境等 13 个指标建立了人工神经网络(ANN)模型,用于对浙江的台风-暴雨灾害风险进行动态预评估<sup>[26]</sup>; OPELLA 等基于卷积神经网络(CNN)和支持向量机(SVM)算法建立了考虑 10 个指标的洪涝灾害风险评估模型<sup>[27]</sup>; LI 等利用 13 个指标对比了逻辑回归、朴素贝叶斯等模型在洪涝风险评估中的表现<sup>[28]</sup>。虽然上述方法在一定程度上提高了评估效率,但是仍存在明显缺点,主要是不易说明各模型参数的作用,类似于“黑箱”操作,无法解释不同指标在灾害风险评估中所起的作用<sup>[29]</sup>。集成学习方法正好能解决上述缺点,该类算法是将多个弱学习器组合建立一个强学习器,能够提高单一模型的泛化能力和稳健性。集成学习不但能提高预测准确率,更重要的是利用其原理可以有效评估指标对最终评估结果的贡献值<sup>[30-32]</sup>。其中, XGBoost 模型被认为是集成学习算法中性能最好的方法之一。

我国目前已形成覆盖国家-省-市-县-乡-村六级的灾情报送网络体系,建成了长时间序列、高精细度的灾害事件案例库,为数据驱动的模型构建提供了保障。本文选取我国南方地区洪涝灾害案例作为训练样本,在县域层级收集可能影响重大洪涝过程灾害风险的 23 项指标,基于 XGBoost 算法建立重大洪涝过程灾害风险评估模型,对南方地区重大洪涝过程中人口、农作物、房屋、经济等损失风险以及综合风险进行评估,并验证指标量和样本量增加对于提升模型性能的意义。

## 1 数据与方法

### 1.1 研究范围

江淮地区、长江中下游地区、华南地区、西南地区等南方地区历来是我国洪涝灾害影响最大和损失最重的区域<sup>[33]</sup>。一般进入主汛期,南方地区会遭受多次大范围降雨过程,基本上每年都会有重大洪涝灾害发生。本文在县级行政单元的尺度上,收集了 2012—2021 年南方地区 650 个有灾情上报的重大洪涝过程案例,除去无法匹配降雨过程的,有 625 个案例用于模型研究,其中, 621 个案例用于模型训练, 4 个案例用于模型测试。训练案例共有 12 640 个县级样本,涉及江苏、浙江、安徽、福建、江西、湖北、湖南、广东、广西、重庆、

四川、贵州、云南等 13 省(自治区、直辖市) 1 120 个县;测试案例是 2021 年 7 月洪涝过程,涉及安徽、湖北、湖南、重庆 4 省(直辖市) 97 个县(图 1)。



图 1 县级训练样本和测试案例分布  
(审图号: GS(2019)1823 号,底图无修改,下同)

### 1.2 评估指标与等级划分

#### 1.2.1 评估指标选取

洪涝灾害是由致灾因子、孕灾环境、承灾体、灾情损失构成的复杂系统,灾情损失是由致灾因子、孕灾环境、承灾体三个子系统相互作用的产物<sup>[34]</sup>。历史灾情能够反映区域灾害风险的大小,数据驱动的风险评估模型是通过探索历史灾情大数据与致灾因子、孕灾环境、承灾体等灾害系统指标的相关关系来建立。本文对上述 625 个洪涝过程分别匹配了 4 个维度 23 项指标(图 2),数据来源如表 1 所示。

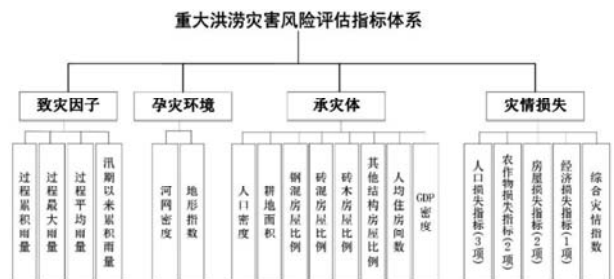


图 2 重大洪涝过程灾害风险评估指标体系

表 1 指标数据来源

数据类型	数据名称	数据时间/年份	数据来源
致灾因子	降雨数据	2012—2021	中国气象局共享数据
孕灾环境	河网、地形数据	2012	应急管理部国家减灾中心孕灾环境数据库
承灾体	人口密度	2019	中国科学院资源环境科学数据中心
	耕地面积	2015	中国科学院资源环境科学数据中心
	房屋结构、人均住房间数	2010	国家统计局第六次人口普查数据
	GDP 密度	2019	中国科学院资源环境科学数据中心
灾情损失	南方地区重大洪涝灾害分县灾情数据	2012—2021	应急管理部国家减灾中心灾情数据库

(1)致灾因子指标。主要是重大洪涝过程的降水数据统计,包括洪涝过程的县域累积雨量、最大雨量、平均雨量以及汛期以来累积降雨量。其中,最大雨量是指洪涝过程的单日最大降雨量;汛期以来累积降雨量统计的是从当年5月1日起到当前过程发生时的累积降雨,考虑的是当年的累积雨量对当前过程的影响。

(2)孕灾环境指标。主要考虑河网和地形影响,包括河网密度和地形指数。其中,河网密度指县域内河道总长度占县域面积的比例,地形指数是水文模拟参数之一,可以表征区域径流面积、坡度等信息。

(3)承灾体指标。主要提供县域承灾体信息,包括人口密度、耕地面积、房屋结构及人均住房间数、GDP密度等指标。

(4)灾情损失指标。主要考虑国家自然灾害统计制度中涉及的指标,包括人口损失、农作物损失、房屋损失和经济损失。其中,人口损失指标用以评估人口损失风险,具体还包括受灾人口、死亡失踪人口、紧急转移安置人口等3项指标;农作物损失指标用以评估农作物损失风险,具体还包括农作物受灾面积、农作物绝收面积等2项指标;房屋损失指标用以评估房屋损失风险,具体还包括倒塌和严重损坏房屋、一般损坏房屋等2项指标;经济损失指标用以评估直接经济损失风险,含1项指标。

另外,综合灾情指数是基于灾情损失各类指标计算的一个指数,用以评估洪涝过程的灾害损失综合风险。综合灾情指数计算有多种方法<sup>[35-37]</sup>,本文采用文献[35]的灾情绝对指数计算方法,对各项指标归一化后,采用加权平均计算,各指标项选取及相应的权重如图3所示。

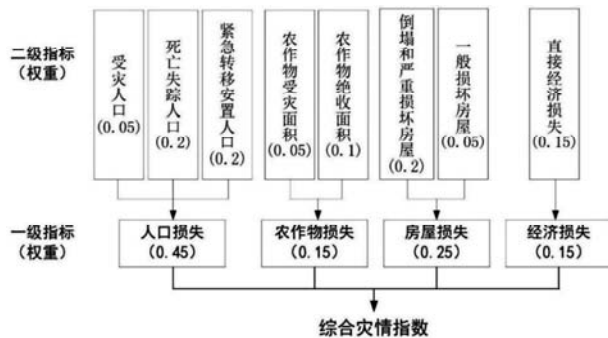


图3 综合灾情指数计算所用的指标及权重

### 1.2.2 灾情损失与风险等级划分

灾情损失与风险等级划分标准统一是评估模型的基础。本文将县域洪涝灾情损失划分为3级:轻灾、中灾、重灾,相对应的县域洪涝过程灾害风

险评估结果等级划分为3级:有一定风险、中风险、高风险。这里认为凡是能预测到有损失可能性的都是有一定风险的,代替传统的低风险等级。实验中将受灾人口、紧急转移安置人口、农作物受灾面积、倒塌和严重损坏房屋、直接经济损失以及综合风险等6项指标分别建模和评估,这些指标常用于实际风险管理工作。考虑到重大洪涝过程风险评估有实际业务倾向,每个指标的灾情损失或风险等级划分阈值的确定主要参考相关灾害管理办法和历史经验(表2)。

### 1.3 XGBoost 算法

XGBoost是集成学习中Boosting家族中的算法<sup>[38]</sup>,目标是将多个弱分类器提升为一个强分类器,用于数据集的分类或回归。XGBoost核心是对算法的损失函数加上了正则化部分,并且损失函数对每一步的误差部分做二阶泰勒展开,更加准确。XGBoost的损失函数形式为:

$$L_t = \sum_{i=1}^m L(y_i, f_{t-1}(x_i) + h_t(x_i)) + \gamma J + \frac{\lambda}{2} \sum_{j=1}^J \omega_j^2. \quad (1)$$

式中: $L$ 表示损失函数, $t$ 表示迭代次数, $i$ 表示第 $i$ 个样本, $m$ 表示样本数量, $f(x)$ 和 $h(x)$ 分别表示强学习器和弱学习器, $J$ 是叶子节点个数, $\omega_j$ 是第 $j$ 个叶子节点的最优值, $\lambda$ 和 $\gamma$ 是正则化系数。XGBoost算法流程总结如下:

模型输入:训练集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ ,  $x_i \in X \subseteq \mathbf{R}^n$ ,  $y_i \in Y \subseteq \mathbf{R}$ ; 损失函数 $L$ ; 最大迭代次数 $T$ 。

模型过程如下:

(1)计算第 $i$ 个样本当前轮损失函数 $L$ 基于 $f_{t-1}(x_i)$ 的一阶导数 $g_{ti}$ 和二阶导数 $h_{ti}$ ,然后计算所有样本的一阶导数和 $G_t$ 以及二阶导数和 $H_t$ 。

(2)尝试分裂决策树,默认分数 $score = 0$ ,  $G$ 和 $H$ 分别为当前需要分裂节点的一阶导数和以及二阶导数和,对指标序号 $k = 1, 2, \dots, K$ ,有

① $G_L = 0, H_L = 0$ 。

②按指标 $k$ 将样本从小到大排列,依次取出第 $i$ 个样本,计算该样本放入左子树后,左右子树的一阶导数和以及二阶导数和:

$$G_L = G_L + g_{ti}, G_R = G - G_L; \quad (2)$$

$$H_L = H_L + h_{ti}, H_R = H - H_L. \quad (3)$$

③更新最大分数:

$$score = \max \left( score, \frac{1}{2} \frac{G_L^2}{H_L + \lambda} + \frac{1}{2} \frac{G_R^2}{H_R + \lambda} - \frac{1}{2} \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \gamma \right). \quad (4)$$

(3)基于最大分数对应的指标分裂决策树。

(4)如果最大得分为0,则当前决策树建立完成,更新当前轮强学习器 $f_t(x)$ ,进入下轮迭代;如果最大得分不为0,则转到步骤(2),继续分裂决策树。

表2 县域灾情损失和风险等级划分阈值

评估指标	轻灾/有一定风险	中灾/中风险	重灾/高风险
受灾人口/人	<20 000	20 000 ~ 80 000	>80 000
紧急转移安置人口/人	<1 000	1 000 ~ 4 000	>4 000
农作物受灾面积/hm <sup>2</sup>	<1 000	1 000 ~ 5 000	>5 000
倒塌和严重损坏房屋/间	<200	200 ~ 1 000	>1 000
直接经济损失/万元	<150	150 ~ 600	>600
综合灾情指数	<0.005	0.005 ~ 0.02	>0.02

#### 1.4 模型建立与性能评估

本文基于 XGBoost 算法分别建立洪涝过程中受灾人口风险、紧急转移安置人口风险、农作物受灾面积风险、倒塌和严重损坏房屋风险、直接经济损失风险、灾害综合风险等 6 个风险评估模型。所有模型以致灾因子、孕灾环境、承灾体三类指标为输入向量, 以不同损失风险等级为输出向量, 具体模型建立流程如图 4 所示。

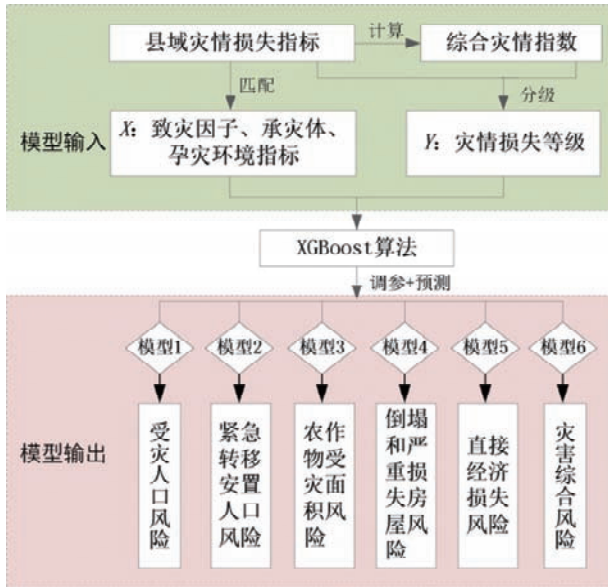


图4 基于 XGBoost 的洪涝灾害风险评估模型建立流程

其中, 模型调参采用 10 折交叉检验优化参数, 对 XGBoost 模型主要 3 个参数进行网格搜索, 分别是弱分类器个数、决策树最大深度和学习率, 6 个模型最优参数及训练集准确率如表 3 所示。

为评估模型精度, 选取准确率( $ACC$ )、查准率( $P$ )、召回率( $R$ )、 $F$  值( $F$ )等 4 个评价指标:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}; \quad (5)$$

$$P = \frac{TP}{TP + FP}; \quad (6)$$

$$R = \frac{TP}{TP + FN}; \quad (7)$$

$$F = \frac{2 \cdot P \cdot R}{P + R}. \quad (8)$$

表3 模型最优参数

评估模型	弱分类器个数	最大深度	学习率	训练集准确率
受灾人口风险	900	4	0.2	0.98
紧急转移安置人口风险	800	2	0.1	0.99
农作物受灾面积风险	1 000	3	0.1	0.93
倒塌和严重损坏房屋风险	800	2	0.1	0.99
直接经济损失风险	900	4	0.2	0.95
灾害综合风险	900	4	0.2	0.94

表4 验证集不同风险等级评估结果

评估指标	准确率	查准率	召回率	$F$ 值
受灾人口风险	0.83	0.89	0.83	0.85
紧急转移安置人口风险	0.97	0.98	0.97	0.97
农作物受灾面积风险	0.83	0.88	0.83	0.85
倒塌和严重损坏房屋风险	0.98	0.98	0.99	0.98
直接经济损失风险	0.63	0.62	0.63	0.61
灾害综合风险	0.84	0.82	0.84	0.82

式中:  $TP$  表示真阳性(True Positive), 即本身是正样本, 预测也是正样本;  $TN$  表示真阴性(True Negative), 即本身是负样本, 预测也是负样本;  $FP$  表示假阳性(False Positive), 即本身是正样本, 预测是负样本;  $FN$  表示假阴性(False Negative), 即本身是负样本, 预测是正样本。上述评价指标中, 准确率表示所有预测正确的样本占总样本的比例; 查准率表示预测为正的样本中是真正的正样本的比例; 召回率表示实际样本中的正例被预测正确的比例;  $F$  值是平衡查准率和召回率的指标, 是二者的调和平均值。

## 2 结果与讨论

### 2.1 案例验证

本文以 2021 年 7 月中旬的南方降雨过程作为案例, 验证基于 XGBoost 算法建立的重大洪涝过程灾害评估模型的应用效果。这次洪涝过程造成安徽、湖北、湖南、重庆等地 97 个县发生灾害, 共造成 90 余万人受灾, 直接经济损失近 10 亿元。

利用建立的模型分别对受灾人口、紧急转移安置人口、农作物受灾面积、倒塌和严重损坏房屋、直接经济损失、灾害综合风险进行评估, 然后与县级实际灾情损失等级对比, 结果如表 4 所示。

可以看出, 模型整体上性能较优, 除直接经济损失风险评估外, 其他风险评估结果的准确率、查准率、召回率和  $F$  值都在 80% 以上, 特别是紧急转移安置人口、倒塌和严重损坏房屋两项的准确率非常高, 达到 97% 和 98%, 这可能因为在测试集中样本不平衡导致, 这两项指标涉及到应急响应启动条件, 该案例中均未达到启动条件; 综合风险评估结果准确率为 84%, 考虑到防止过拟合的参数设定, 该模型精度达到了预期效果, 通过实际灾情(图 5a)与风险评估结果(图 5b)的空间对比, 可以看出, 模型评估在安徽、湖北、湖南、重庆都能够较好地预测洪涝过程的灾害风险; 直接经济损失风险评估结果准确率相对较低, 只有 63%, 可能的原因是直接经济损失在实际统计中较为复杂, 包含了很多统计子项, 不同地区、不同时间的直接经济损失统计标准可能存在较大差异。



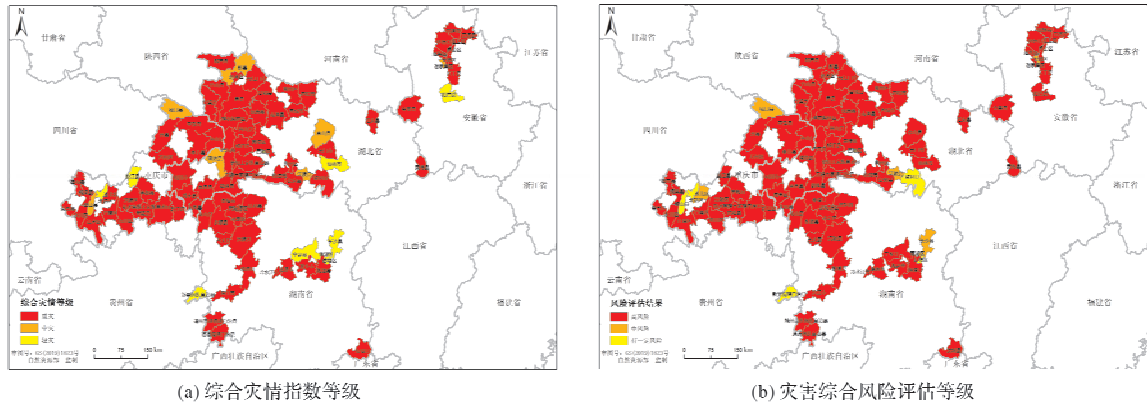


图5 测试集案例中实际综合灾情指数与模型评估的灾害综合风险对比

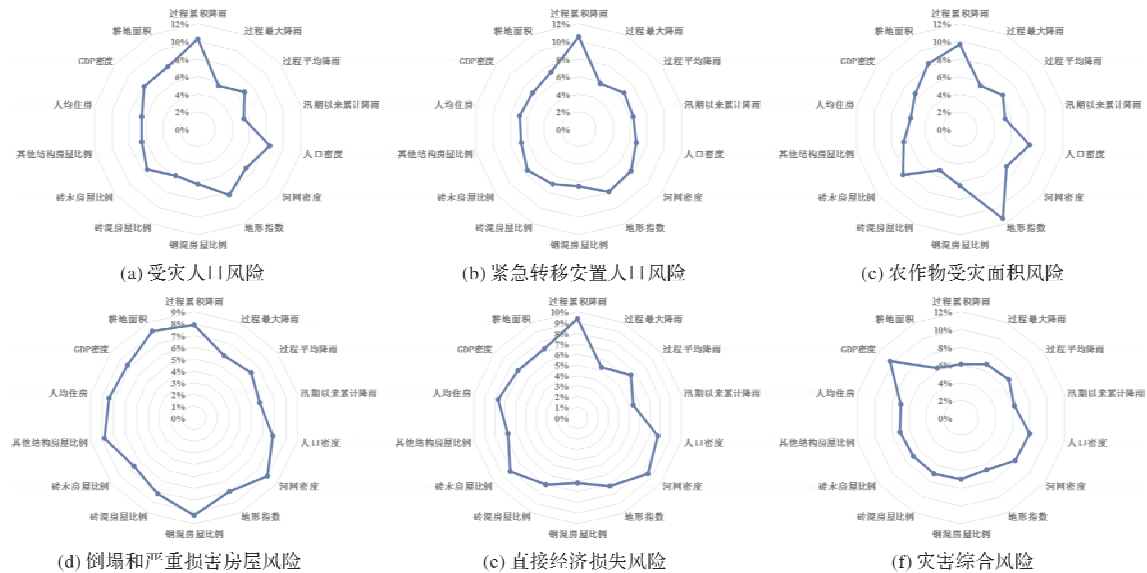


图6 不同风险评估类型的指标重要性

## 2.2 指标重要性

要理解影响评估结果的各种因素,有必要对各评估指标的具体贡献进行计算。XGBoost 算法计算指标重要性的原理是一个输入指标在决策树分枝时是被选取为分枝特征次数越多,则该特征越重要。本文分别对6类风险评估目标进行了指标重要性计算,结果如图6所示。其中,影响受灾人口风险评估最重要的指标是过程累积降雨、人口密度、地形指数,表明致灾因子、承灾体、孕灾环境都对评估结果有所贡献;影响紧急转移安置人口风险评估最重要的指标是过程累积降雨、地形指数、河网密度,这与实际灾害应对过程相符,一般雨量大、地势险、离水近的人口在灾害应急中容易被转移;农作物受灾面积风险与地形指数、过程累积降雨、耕地面积、砖木房屋比例、人口密度有较大关系,特别是地形指数影响突出,说明农作物的受灾与孕灾环境有很强的关联;倒塌和严重损坏房屋风险明显受过程累积降雨、人均住房数、房屋结构比例、河网密度等因素影响,承灾体指标重要性突出;直接经济损失风险影响因素主要是过程累积降雨、河网密度,可见直接经济损失大小与致灾因子、孕灾环境密不可分;灾害综合风险的主要影响因素是GDP密度,可能的原因是GDP本身就是综合性指标,GDP既能反

映区域的承灾体综合暴露度,也在一定程度上反映了区域的承灾体脆弱性,即一般情况下可以认为GDP高的地区比GDP低的地区综合防灾减灾能力要更强。总体来说,不同指标对不同风险评估结果的贡献并不完全相同,也没有一个指标的贡献率可以达到忽略不计的程度,各指标贡献率都在5%~12%之间。

## 2.3 指标量和样本量对评估结果的影响

为考察指标数量对评估模型精度的影响,本文对不同维度的输入指标进行了组合,比较了仅用致灾因子、致灾因子和孕灾环境组合、致灾因子和承灾体组合以及采用全部指标在洪涝灾害风险评估结果中的准确率(表5)。通过对比发现,指标量的变化对受灾人口风险、紧急转移安置人口风险、倒塌和严重损坏房屋风险3个模型评估结果影响较小。然而,指标量对农作物受灾面积风险、直接经济损失风险和灾害综合风险3个模型评估准确率有较大影响,如果模型输入只有致灾因子,准确率最低,比全指标偏低10%~15%;在致灾因子基础上,无论加入孕灾环境和承灾体指标,对准确率都有较大提升,而且承灾体指标比孕灾环境指标提升多,因为承灾体指标子项更多;而将所有指标一同作为输入,准确率最高,说明指标量对评估结果有很大影响。

表 5 不同指标量的风险评估准确率

指标量	致灾因子	致灾因子 + 孕灾环境	致灾因子 + 承灾体	致灾因子 + 孕灾环境 + 承灾体
受灾人口风险评估准确率/%	80	82	82	83
紧急转移安置人口风险评估准确率/%	94	97	96	97
农作物受灾面积风险评估准确率/%	73	78	80	83
倒塌和严重损坏房屋风险评估准确率/%	94	95	97	98
直接经济损失风险评估准确率/%	48	51	59	63
灾害综合风险评估准确率/%	72	79	82	84

一般对某个区域进行灾害风险评估, 往往仅选用该区域内相关灾害案例进行分析<sup>[19,21]</sup>, 这样可能导致区域样本量较少。为考察样本数量对评估模型精度的影响, 本文对验证集中的安徽、湖北、湖南和重庆 4 个省(直辖市)案例进行对比。首先, 提取各自省份(直辖市)2012—2021 年的县域灾害数据样本; 然后利用各省(直辖市)样本分别进行训练, 建立各省(直辖市)基于单省份样本的综合风险评估模型; 最后比较基于单省份样本量的模型和基于全样本量的模型在洪涝灾害综合风险评估结果中的准确率(表 6)。可以看出, 全样本评估对湖南评估结果准确率提升了 13%, 其他省份也有 5%~10% 的提升, 这说明样本量的增加对各省评估准确率都有较大增益。

表 6 不同样本量的灾害综合风险评估准确率

省份	安徽	湖北	湖南	重庆
2012—2021 年单省份样本量	366	1 258	1 769	798
单省份样本准确率/%	83	81	65	77
2012—2021 年全样本量	12 640	12 640	12 640	12 640
全样本准确率/%	92	86	78	86

### 3 结论

本文基于我国南方地区 625 个重大洪涝过程案例的将近 30 万条指标, 利用 XGBoost 算法建立了重大洪涝过程灾害评估模型, 并以 2021 年 7 月中旬南方地区洪涝过程数据对模型进行了验证。结果表明该模型可用于重大洪涝灾害事件发生前对受灾人口风险、紧急转移安置人口风险、农作物受灾面积风险、倒塌和严重损坏房屋风险、灾害综合风险等 5 个模型在风险评估中的准确率、查准率、召回率和  $F$  值等性能指标均在 80% 以上, 说明模型有较好的泛化性能, 能够用于实际灾害风险评估工作中。

(1) 基于 XGBoost 算法建立的重大洪涝过程灾害风险评估模型适用性较好。通过测试案例验证, 受灾人口风险、转移安置人口风险、农作物受灾面积风险、倒塌和严重损坏房屋风险、灾害综合风险等 5 个模型在风险评估中的准确率、查准率、召回率和  $F$  值等性能指标均在 80% 以上, 说明模型有较好的泛化性能, 能够用于实际灾害风险评估工作中。

(2) 模型可以通过计算重要性评价指标对风险评估结果的贡献度。除过程累积降雨指标对大部分评估目标都有影响外, 不同风险评估目标的影响因素不同, 如受灾人口、倒塌和严重损坏房屋受承灾体因素影响较大, 紧急转移安置人口、农作物受灾面积、直接经济损失主要受孕灾环境影响, 而灾害综合风险主要影响因素是 GDP 密度。指标重要性增加了风险评估模型的可解释性, 提升了指标与评估结果之间关系的理解, 有助于改

进对机器学习算法“黑箱”模式的认知。

(3) 指标量和样本量对于数据驱动的评估模型有重要作用。集成学习算法在灾害风险评估中淡化了危险性、脆弱性等灾害机理, 纯粹利用灾害系统相关数据进行学习, 建模较为简单, 这也要求评估指标和样本数量要有足够积累。一方面, 致灾因子指标、孕灾环境指标、承灾体指标都对灾害风险评估结果有重要影响, 利用全指标量比仅用致灾因子指标可以提高 10%~15% 的评估准确率; 另一方面, 样本数量增加 1~2 个数量级能够提高 5%~13% 的评估准确率。这表明灾害大数据对于提升灾害风险评估模型性能有很大帮助。

(4) 模型仍存在一定的不确定性。由于选择的训练案例源自全国各地上报灾情, 灾情本身是人为统计, 存在着时间差异和地区差异, 这可能也是模型在直接经济损失风险评估中没有其他风险评估结果准确性高的原因之一。另外, 样本本身偏向于重大洪涝灾害过程, 对于一般灾害的评估效果尚待验证。

本文利用灾害大数据建立了县域洪涝过程灾害风险评估模型。随着社会经济的快速发展, 区域的承灾体和孕灾环境会发生很多变化, 在今后的研究中, 需要不断引入最新数据, 更新和累积大数据, 提升模型的可靠性。总结下一步工作, 重点有三个方向: 一是继续完善指标体系和样本分布, 利用第一次全国自然灾害综合风险普查数据对指标进行更新, 进一步完善模型; 二是收集北方地区重大灾害过程案例, 并验证模型在北方地区是否具有通用性; 三是模型的推广应用, 比如用于确定针对历史灾情分析时各指标权重, 或用于缺乏资料地区的洪涝灾害风险评估等。

### 参考文献:

- [1] Ministry of Emergency Management of the People's Republic of China [EB/OL]. [2022-04-01]. [https://www.mem.gov.cn/xw/yjglbgzdt/202201/t20220123\\_407204.shtml](https://www.mem.gov.cn/xw/yjglbgzdt/202201/t20220123_407204.shtml).
- [2] CRED, UNDRR. The human cost of disasters: an overview of the last 20 years (2000—2019) [M]. Geneva: UNDRR, 2020.
- [3] IPCC. Climate change 2021: the physical science basis [M/OL]. [2022-03-01]. [https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC\\_AR6\\_WGI\\_Full\\_Report.pdf](https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Full_Report.pdf).
- [4] WEIS S, AGOSTINI V N, ROTH L M, et al. Assessing vulnerability: an integrated approach for mapping adaptive capacity, sensitivity, and exposure [J]. Climatic Change, 2016, 136 (3): 615—629.
- [5] ZHANG J, CHEN Y. Risk assessment of flood disaster induced by typhoon rainstorms in Guangdong Province, China [J]. Sustainability, 2019, 11(10): 2738.
- [6] 黄崇福, 郭君, 艾福利, 等. 洪涝灾害风险分析的基本范式及其应用 [J]. 自然灾害学报, 2013, 22(4): 11—23.
- [7] 徐伟, 田玉刚, 张勇, 等. 综合风险防范: 长江三角洲地区自然致灾因子与风险等级评估 [M]. 北京: 科学出版社, 2014.

- [8] 卞洁. 长江中下游暴雨洪涝灾害的风险性评估与预估[D]. 南京: 南京信息工程大学, 2011.
- [9] APEL H, O. MARTÍNEZ T, HUNG N, et al. Combined fluvial and pluvial urban flood hazard analysis: Concept development and application to Can Tho city, Mekong Delta, Vietnam[J]. *Natural hazards and Earth System Sciences*, 2016, 16(4): 941–961.
- [10] 杜晓燕, 黄岁樑, 赵庆香. 基于信息扩散理论的天津旱涝灾害危险性评估[J]. *灾害学*, 2009, 24(1): 22–25.
- [11] 张丽娟, 李文亮, 张冬有. 基于信息扩散理论的气象灾害风险评估方法[J]. *地理科学*, 2009, 29(2): 250–254.
- [12] FEMA. Hazus Multi – Hazard Loss Estimation Methodology: Earthquake Model (Hazus – MH. Technical Manual 21) [M]. Washington, DC: Mitigation Division, Department of Homeland Security, Federal Emergency Management Agency, 2012.
- [13] ERN – AL. CAPRA [EB/OL]. [2022 – 04 – 01]. <https://ecapra.org/topics/risk-assessment>.
- [14] KRON W. Flood risk = hazard × exposure × vulnerability[M]. New York: Science Press, 2002.
- [15] 黄崇福. 自然灾害动态风险分析基本原理的探讨[J]. *灾害学*, 2015, 30(2): 1–7.
- [16] 邹铭, 范一大, 杨思全, 等. 自然灾害风险管理与预警体系[M]. 北京: 科学出版社, 2010.
- [17] GUO G, LIU L, LI Y, et al. A Vulnerability curve method to assess risks of climate – Related hazards at county level[J]. *Atmosphere*, 2021, 12(8): 1022.
- [18] 高歌, 王长源, 吴先华, 等. 山洪沟暴雨洪涝灾害情境下居民室内财产的脆弱性曲线研究: 以2016年9月福建省长泰县马洋溪山洪灾害为例[J]. *灾害学*, 2018, 33(3): 126–131.
- [19] 杨佩国, 靳京, 赵东升, 等. 基于历史暴雨洪涝灾情数据的城市脆弱性定量研究: 以北京市为例[J]. *地理科学*, 2016, 36(5): 733–741.
- [20] 尹卫霞, 余瀚, 崔淑娟, 等. 暴雨洪水灾害人口损失评估方法研究进展[J]. *地理科学进展*, 2016, 35(2): 148–158.
- [21] 袁祝香, 张晨琛, 吕忠, 等. 吉林省重大暴雨过程灾害损失风险预评估[J]. *冰川冻土*, 2016, 38(2): 395–401.
- [22] 叶金玉, 林广发, 张明锋. 自然灾害风险评估研究进展[J]. *防灾科技学院学报*, 2010, 12(3): 20–25.
- [23] 徐伟, 杜鹏, 乔宇, 等. 亚太经济合作组织灾害风险评估: 典型模型和研发建议[J]. *中国减灾*, 2021(21): 16–21.
- [24] 段华明, 何阳. 大数据对于灾害评估的建构性提升[J]. *灾害学*, 2016, 31(1): 188–192.
- [25] 何昕宇, 田文翀, 张智宇, 等. 基于数据驱动的洪涝风险评估方法研究进展[J]. *人民珠江*, 2022, 43(5): 60–67.
- [26] 刘芳. 洪涝灾害风险评估方法研究: 以浙江省台风暴雨为例[D]. 杭州: 浙江大学, 2021.
- [27] OPELLA J, HERNANDEZ A A. Developing a Flood Risk Assessment Using Support Vector Machine and Convolutional Neural Network: A Conceptual Framework [C]//2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA). IEEE, 2019: 260–265.
- [28] LI X, YAN D, WANG K, et al. Flood Risk Assessment of Global Watersheds Based on Multiple Machine Learning Models [J]. *Water*, 2019, 11(8): 1654.
- [29] WANG Z, LAI C, CHEN X, et al. Flood hazard risk assessment model based on random forest [J]. *Journal of Hydrology*, 2015, 527: 1130–1141.
- [30] MERZ B, KREIBICH H, LALL U. Multi – variate flood damage assessment: a tree – based data – mining approach [J]. *Natural hazards and earth system sciences*, 2013, 13(1): 53–64.
- [31] 林森, 郭桂祯, 刘蓓蓓. 基于梯度提升决策树 (GBDT) 算法的南方洪涝灾害房屋倒塌评估模型[J]. *灾害学*, 2020, 35(1): 109–112, 130.
- [32] 佟金萍, 张涵玥, 刘辉, 等. 基于 XGBoost 的长三角核心城市内涝风险评估及影响因素分析[J]. *水利水电技术 (中英文)*, 2021, 52(10): 1–11.
- [33] 万金红, 张葆蔚, 刘建刚, 等. 1950–2013 年我国洪涝灾情时空特征分析[J]. *灾害学*, 2016, 31(2): 63–68.
- [34] 史培军. 灾害研究的理论与实践[J]. *南京大学学报*, 1991, 11: 37–42.
- [35] 袁艺. 2000–2007 年省级区域自然灾害灾情分析[J]. *自然灾害学报*, 2011, 20(1): 156–162.
- [36] 廖永丰, 赵飞, 王志强, 等. 2000–2011 年中国自然灾害灾情空间分布格局分析[J]. *灾害学*, 2013, 28(4): 55–60.
- [37] 张鹏, 张云霞, 孙舟, 等. 综合灾情指数: 一种自然灾害损失的量化评价方法[J]. *灾害学*, 2015, 30(4): 74–78.
- [38] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system [C]//Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining. ACM, 2016: 785–794.

## Disaster Risk Assessment at County Level of a Heavy Flooding Driven by Disaster Big Data

LIN Sen, LIU Beibei, YAN Xue, SUN Ning, GUO Guizhen

(National Disaster Reduction Center of China, Ministry of Emergence Management Department of China,  
Beijing 100124, China)

**Abstract:** Disaster risk assessment is a key step in the risk management of a heavy flooding. The data – driven machine learning models for risk assessment is not only convenient for modeling, but also can comprehensively consider various indices of the flood disaster system including disaster – causing factors, disaster pregnant environment, disaster – bearing body and disaster losses. The ensemble methods represented by XGBoost can calculate each index importance and thus improve the interpretability of the model. Based on 625 heavy flooding cases in southern China, we collected nearly 300, 000 data of 23 county – level indices and established 6 models using the XGBoost algorithm, which can be used before the occurrence of a heavy flooding to assess the risk of the affected population, the population to be transferred and resettled, the affected crop area, the collapse and serious damage houses, the direct economic loss and the comprehensive risk of the flood disaster. The models are verified by a case in 2021 and the overall accuracy of disaster risk assessment exceeded 80%, which proves that the models in this study have a good generalization ability and can be utilized in practical disaster assessment. Compared with the results only using the disaster causing factor index, utilizing the whole index can boost the evaluation accuracy by 10–15%. Moreover, increasing the sample size by 1–2 orders of magnitude can improve the accuracy by 5%–13%. This indicates that the accumulation of disaster big data will be of great significance to disaster risk assessment.

**Keywords:** flood disaster; big data; risk assessment; evaluation index; XGBoost