

黄星, 周鸿娇. 暴雨灾害下长输管道风险因子的双阶皮尔逊筛选方法[J]. 安全与环境学报, 2026, 26(6): 2417–2431.

HUANG X, ZHOU H J. Two-stage Pearson correlation screening method for identifying risk factors in long-distance pipelines during rainstorm disasters[J]. Journal of Safety and Environment, 2026, 26(6): 2417–2431.

暴雨灾害下长输管道风险因子的双阶皮尔逊筛选方法*

黄星¹, 周鸿娇²

(1 西南石油大学经济管理学院, 成都 610500; 2 西南石油大学土木工程与测绘学院, 成都 610500)

摘要: 暴雨天气下构建科学有效的长输管道受灾风险因子体系, 对风险诊断与预测具有重要意义。针对相关研究的不足, 设计了双阶皮尔逊关联的风险因子筛选方法。在初选阶段, 采用皮尔逊相关矩阵和变异系数双重准则, 降低候选风险因子间信息冗余, 形成初选风险因子体系; 在终选阶段, 设计信息冗余贡献度算法, 循环计算初选风险因子的信息冗余贡献度指数, 筛选对整体相关性贡献较小的风险因子, 构建终选风险因子体系。为验证方法的有效性, 选取 K -中心和 K -均值聚类算法对比。结果表明: 所提方法的风险因子信息解释强度为 1.033, 累计信息贡献率达 89.9%, 显著高于 K -中心聚类算法 (1.007, 51.2%) 和 K -均值聚类算法 (1.057, 53.7%)。该方法可为暴雨天气下长输管道风险预测与预警提供有力支持。

关键词: 安全系统学; 暴雨灾害; 长输管道; 风险因子筛选; 双阶皮尔逊关联法

中图分类号: X43; TE88

文献标志码: A

文章编号: 1009-6094(2026)06-2417-15

DOI: 10.13637/j.issn.1009-6094.2025.1057

0 引言

长输管道作为能源输送命脉, 在暴雨天气频发环境下承受着生态破坏、经济损失及公共安全的三重危机。暴雨天气下长输管道受灾影响巨大, 如 2020 年 7 月 14 日贵州遵义桐梓县境内的输油管道因暴雨引发滑坡导致柴油泄漏, 造成严重的经济损失和环境污染。因此, 强化暴雨天气下长输管道风险管理, 科学实施灾害风险评估与防控, 具有重要实践意义。然而, 针对暴雨天气下长输管道受灾风险因子构建和筛选, 仍存在诸多不足, 亟待深入研究。

暴雨天气下长输管道受灾风险因子相关研究主要聚焦于风险因子识别维度与筛选方法上。在识别维度方面, 研究者们主要从地形、地质、气象、管道状况及人文社会因素等维度, 将风险因子归纳为灾害发生风险概率、长输管道易损性及受灾风险后果三个方面^[1-3]。2015 年, 钟威等^[4]首次系统性地建立了包含地理特征、气象水文与地质灾害等风险因子的长输管道评价体系, 为后续相关研究奠定了基础。Badida 等^[5]和李越等^[6]则在此基础上构建了更加完善的风险因子评价体系。晋良海等^[7]聚焦管道自身脆弱性, 构建了包含管道易损性、管道受灾后果等维度的风险因子评价体系。此外, 部分研究聚焦复合风险来源, 如冼国栋等^[8]基于地貌特征构建了包

含地质灾害危险性和管道易损性的风险因子评价体系。Haile 等^[9]从第三方活动、结构性缺陷、管理漏洞等维度建立了长输管道风险因子评价体系。当前风险因子筛选方法主要分为信息冗余筛选和信息贡献度筛选两类。在信息冗余筛选中, 学者们从局部、整体和综合视角构建风险因子体系。局部视角强调风险因子间的相关性, 如王灿星等^[10]采用 Pearson 相关系数进行风险因子局部信息冗余筛选。Zahid 等^[11]和 Pan 等^[12]通过 K -中心点聚类、 K -均值聚类算法实现了风险因子局部冗余信息剔除。整体视角关注全局信息冗余控制, 如陈洪海^[13]提出病态循环指数法系统消除风险因子体系的整体信息冗余。综合视角从局部和整体两个方面进行信息冗余剔除。2005 年, Peng 等^[14]首次提出最小冗余最大相关的特征选择方法, 为基于局部相关性与整体冗余筛选风险因子提供了思路。Wu 等^[15]构建了局部-整体联合筛选框架以获得城市生态环境脆弱性因子。信息贡献度筛选大多通过主成分分析、变异系数、信息熵及层次分析法实现。刘威等^[16]结合信息熵筛选管道风险因子。熊俊楠等^[17]结合模糊综合评价和信息熵筛选泥石流灾害下长输管道的风险因子。侯志强等^[18]采用层次分析法筛选管道安全风险因子。

文献分析表明, 现有研究不足主要体现为: 其一, 当前风险因子体系未覆盖暴雨下灾害全过程, 忽

* 收稿日期: 2025-06-30

作者简介: 黄星, 教授, 博士/后, 博士生导师, 从事网络舆情与应急管理研究, huangxing6213@126.com。

基金项目: 教育部人文社科研究基金规划项目(22YJA630031); 四川省哲学社会科学规划重点项目(SC24ST005)

略了灾害监测、管道管理与防护、应急处置等关键环节,容易造成高危风险点漏判,削弱预警准确性、应急决策针对性与应急措施及时性;其二,现有方法对信息冗余筛选的视角不全面,局限于局部或整体,易遗漏冗余信息,致使风险因子信息冗余度筛选不佳,影响风险预警效率和精度;其三,部分方法筛选过度,信息严重丢失,导致风险预警失效或应急响应滞后。

基于以上分析,本文根据灾害链理论,从灾害发生、发展至控制全过程,构建包含致灾因子、受灾体、抗灾体和外部环境维度的暴雨天气下长输管道受灾风险因子体系。基于皮尔逊相关系数设计双阶皮尔逊关联法,从局部和整体两个角度对风险因子进行全面信息冗余筛选。首先,通过相关系数与变异系数剔除因子间的局部信息冗余,再结合信息冗余贡献指数识别并剔除体系内的整体信息冗余,使筛选结果既覆盖完整灾害链,又能确保因子的独立性和代表性。同时,设定累计信息贡献率控制筛选强度,避免过度筛选造成信息缺失,确保因子体系在精简后的解释力。筛选后的风险因子体系在降低因子间信息冗余、减少强相关干扰的基础上,可作为后续开展权重分配与风险评价模型构建的前提。

1 研究设计

基于灾害链理论,结合学术文献、管道安全规范及暴雨天气下长输管道事故调查报告等多源信息,系统识别候选风险因子库。在此基础上,通过专家打分反映实践经验下的候选风险因子重要性。但由于部分候选风险因子间得分区分度不足,难以形成清晰的优先序列,直接采用则易造成体系臃肿、结构重复、解释效率下降,因此仅作为初步参考。为此,本文在专家判断基础上引入客观筛选机制,设计双阶皮尔逊关联法对候选风险因子体系进行优化,其过程分为两个阶段(见图1)。

(1)初选阶段:计算候选风险因子相关系数矩阵。随机比对两个风险因子,当两者相关系数超过设定阈值时,剔除变异系数较小者,保留信息量更大的因子;若未超过阈值,则两者均予以保留。通过不断迭代,形成初选风险因子体系。此阶段的目的在于剔除因子间的局部冗余,确保因子间具有较强的独立性与代表性。

(2)终选阶段:基于初选结果,计算剔除单个因子后的整体相关度与每个因子的信息冗余贡献指数,结合累计信息贡献率要求,剔除贡献较大的因

子,最终形成终选风险因子体系。此阶段的目的在于剔除风险因子体系的整体冗余,确保因子体系既精简又具有较强解释力。

完成筛选后,进一步通过案例验证和稳定性检验对方法进行检验,并将筛选结果与 K -中心点聚类法和 K -均值聚类算法进行对比,以确认双阶皮尔逊关联法的可靠性和适用性。相较于传统仅依赖两两相关性的皮尔逊相关系数法,双阶皮尔逊关联法的优势主要体现在三个方面:一是构建局部与整体结合的筛选框架,避免单一角度分析造成的遗漏;二是引入变异系数作为判别准则,使保留因子更具信息量和代表性;三是计算信息冗余贡献指数,提升因子体系的独立性和解释力。

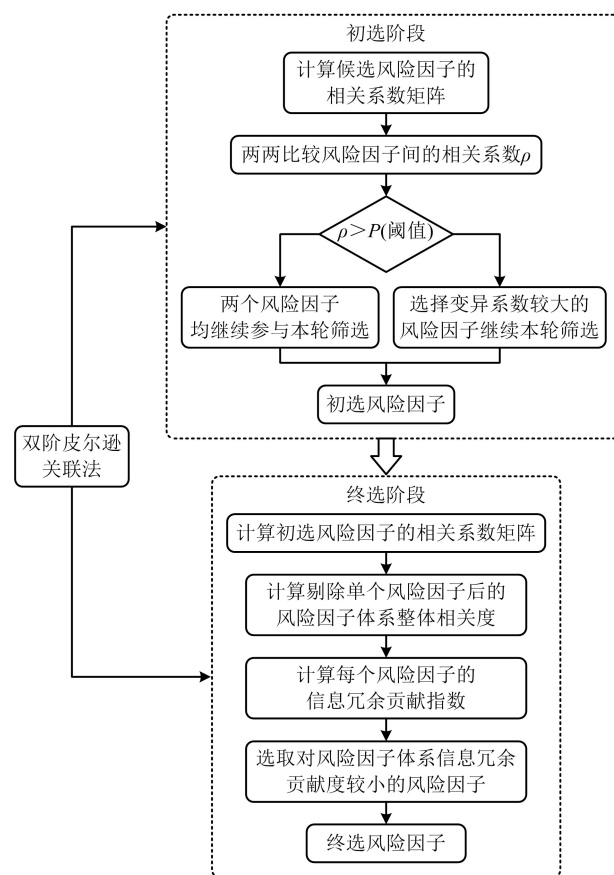


图1 双阶皮尔逊关联法筛选流程图

Fig. 1 Flowchart of two-stage Pearson correlation method for screening

2 风险因子筛选过程

2.1 风险因子候选

灾害链理论认为,灾害是由致灾因子、受灾体、抗灾体三类内部因子,以及自然、社会环境两类外部因子共同构成的系统^[19]。其演化过程本质上是致

灾因子链式传递、受灾体脆弱性暴露、抗灾体干预和环境调控共同作用的动态系统。内部因子与外部环境交互,最终决定灾害风险水平。致灾因子经外部环境作用于受灾体,破坏力受抗灾体制约;抗灾体通过干预削弱灾害链传递,效果受外部环境影响;自然环境调控传导路径,社会环境则影响灾害后果规模。

本文基于灾害链理论,构建包含致灾因子、受灾体、抗灾体及自然与社会环境的暴雨天气下长输管道受灾候选风险因子框架(见图2)。其中,致灾因子为暴雨及暴雨引起的次生灾害,受灾体为长输管道,抗灾体为管道防灾应急所需的人员、物资与信息资源,外部环境包括管道周边自然与社会环境。暴雨及次生灾害经自然环境作用于管道,抗灾减灾措施如灾害监测、管道管理与抗灾防护、应急响应等可削弱其破坏力,抗灾减灾效能受社会环境支撑,社会环境亦影响灾害后果。各类因子协同作用,共同决定管道受灾风险,形成多维影响机制。

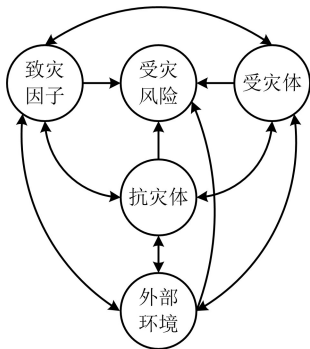


图2 暴雨天气下长输管道受灾风险因子框架

Fig.2 Risk factor framework for long-distance pipeline disaster under heavy rainfall

根据以上分析,在致灾因子方面提出“暴雨气象特征”维度的风险因子^[20],在受灾体方面提出“管道自身情况”维度的风险因子^[21],在抗灾体方面提出“气象监测与预警水平”“管道防护措施情况”“应急响应水平”“管道管理水平”维度的风险因子^[22],并根据自然环境与社会环境两类外部环境,提出“地形地貌条件”“地质条件”“管道周边社会环境”维度的风险因子。候选风险因子体系见表1。

表1中,针对各维度风险因子的选取依据和含义解释如下。

(1)暴雨气象特征维度。参考 GB/T 33669—2017《极端降水监测指标》等相关国家标准,选取“降雨量”“降雨持续时间”“降雨强度”“小时最大降雨量”“降雨频率”风险因子。考虑短历时降水的

表1 候选风险因子体系
Table 1 Candidate risk factor system

一级维度	二级维度	风险因子	风险因子符号
致灾因子	暴雨气象特征	降雨量	X_1
		降雨强度	X_2
		降雨持续时间	X_3
		降雨频率	X_4
		小时最大降雨量	X_5
		短历时强降雨频次	X_6
		暴雨发生的季节	X_7
自然环境	地形地貌条件	地形坡度	X_8
		植被覆盖度	X_9
		水体分布	X_{10}
	地质条件	沟谷分布	X_{11}
		土壤类型	X_{12}
		断层分布	X_{13}
		地下水位	X_{14}
受灾体	管道自身情况	岩性	X_{15}
		管道埋深	X_{16}
		管道壁厚	X_{17}
		管道直径	X_{18}
		管道材质与抗压强度	X_{19}
		管道服役年限	X_{20}
		管道防腐层状况	X_{21}
		管道运行内压	X_{22}
		管道敷设方式	X_{23}
		管道周边社会环境	X_{24}
抗灾体	气象监测与预警水平	周边交通设施分布	X_{25}
		人口密度	X_{26}
		气象监测站点覆盖密度	X_{27}
	管道防护措施情况	暴雨预警准确率	X_{28}
		预警信息传递效率	X_{29}
		排水设施建设情况	X_{30}
	应急响应水平	水土保持措施情况	X_{31}
		抗灾加固措施情况	X_{32}
		应急物资储备情况	X_{33}
	管道管理水平	应急演练频率与效果	X_{34}
		与周边单位应急联动能力	X_{35}
		巡检频率与覆盖范围	X_{36}
	社会环境	管理人员培训与专业能力	X_{37}
		信息化管理水平	X_{38}

致灾特性和暴雨季节性变化特征,补充“短历时强降雨频次”“暴雨发生的季节”风险因子^[23-24]。

(2) 地形地貌与地质条件维度。参考 GB/T 40112—2021《地质灾害危险性评估规范》,选取“地形坡度”“水体分布”“沟谷分布”“土壤类型”“断层分布”“地下水位”“岩性”风险因子。植被具有固土和截留雨水作用,补充选取“植被覆盖度”风险因子^[25]。其中,“降雨强度”表征一定时段内的平均降水速率,反映持续性降雨对土体浸润和边坡稳定的长期影响;“小时最大降雨量”突出极端短时暴雨的峰值,强调在极短时间内产生的强烈径流可能导致排水系统超载,并诱发沟谷型泥石流、管道裸露段冲刷破坏等突发灾害。

(3) 管道自身情况维度、管道防护措施情况维度。依据 GB/T 50253—2014《输油管道工程设计规范》,选取“管道埋深”“管道壁厚”“管道直径”“管道材质与抗压强度”“管道服役年限”“管道防腐层状况”“管道敷设方式”“排水设施建设情况”“水土保持措施情况”“抗灾加固措施情况”“管道运行内压”风险因子^[26]。其中,“管道敷设方式”不仅决定管道所处的地质与水文条件,也直接反映道路、河流等荷载环境对管道的作用,在暴雨下会影响管道的整体风险水平及灾后抢修难度。

(4) 管道周边环境、管道管理水平维度、应急响应水平维度。参考 GB/T 32167—2015《油气输送管道完整性管理规范》,选取“周边交通设施分布”“人口密度”“巡检频率与覆盖范围”“管理人员培训与专业能力”“信息化管理水平”“应急物资储备情况”“应急演练频率与效果”“与周边单位应急联动能力”风险因子。考虑到“应急物资储备情况”“应急演练频率与效果”“与周边单位应急联动能力”“管理人员培训与专业能力”在直接量化上存在难度,其量化方式将在后续风险评估阶段确定。其中,应急物资储备根据关键物资种类齐全度、数量充足度和可用性进行分级,应急演练根据年度开展频次与参与范围分级,应急联动能力由交通可达性与地形条件综合判定,管理人员能力按人均年度培训学时与培训考核合格率分级。

(5) 气象监测与预警水平维度。参考 DB36/T 1984—2024《气象灾害预警信息发布质量管理规范》、QX/T 469—2018《气候可行性论证规范》,选取“预警信息传递效率”“暴雨预警准确率”“气象监测站点覆盖密度”风险因子。

2.2 数据获取

基于前期研究,系统识别出暴雨天气下长输管道受灾的 37 个候选风险因子。为评估其重要性,设

计并发放调查问卷,邀请相关领域专家进行打分,作为因子体系构建的依据。问卷采用李克特 5 级量表(1 分为非常不重要,2 分为比较不重要,3 分为一般重要,4 分为比较重要,5 分为非常重要),量表情况见表 2。调研采用分层抽样,共发放 280 份,回收有效问卷 212 份(有效率为 75.7%),覆盖高校、政府与企业。受访者背景涵盖气象、地质、石油、环境与安全等领域,95% 以上具有交叉研究经验,工作年限分布均衡,样本具有较强代表性。不同专业关注点存在差异,自然类关注环境条件,工程与管理类重视管道状态与应急管理,体现了风险因子体系构建需综合多学科视角。问卷设置必填项,异常值经离散度检验与人工复核剔除,保证数据完整可靠。

为确保调查结果的有效性,对问卷数据进行了信度与效度检验。在信度方面,克隆巴赫系数(Cronbach's α)结果显示总体量表信度为 0.921,各维度信度均大于 0.7,收集的数据符合统计要求。效度分为内容效度和结构效度。问卷基于国内外成熟量表框架设计,并经石油工程、安全管理、地质工程等领域 5 名专家修订,保证内容效度。结构效度经凯泽-梅耶-奥尔金(Kaiser-Meyer-Olkin, KMO)取样适配性检验和巴特利特(Bartlett)球形度检验加以验证。整体 KMO 值为 0.889,大部分维度 KMO 值高于 0.7。尽管部分维度 KMO 值在 0.500 ~ 0.697 之间,但其因子荷载系数均在 0.526 ~ 0.898 之间,显著高于 0.4,表明各因子能有效反映所处维度特征。所有维度显著性水平均达 $p < 0.001$,满足统计要求,整体结构效度良好,结果见表 3。

表 2 量表情况与信度检验结果

Table 2 Scale situation and reliability test results

二级维度	风险因子数量	Cronbach's α	总体 Cronbach's α
暴雨气象特征	7	0.877	0.921
地形地貌条件	4	0.813	
地质条件	4	0.804	
管道自身情况	8	0.897	
管道周边社会环境	2	0.760	
气象监测与预警水平	3	0.783	
管道防护措施情况	3	0.765	
应急响应水平	3	0.729	
管道管理水平	3	0.765	

表 3 结构效度检验结果

Table 3 Results of structural validity tests

二级维度	题项	KMO	Bartlett 球形度检验	自由度	显著性	因子荷载
暴雨气象特征	7	0.919	601.689	21	<0.001	0.725 ~ 0.780
地形地貌条件	4	0.799	272.272	6	<0.001	0.749 ~ 0.840
地质条件	4	0.789	259.075	6	<0.001	0.768 ~ 0.830
管道自身情况	8	0.926	788.929	28	<0.001	0.738 ~ 0.799
管道周边社会环境	2	0.500	98.877	1	<0.001	0.898
气象监测与预警水平	3	0.701	178.976	3	<0.001	0.815 ~ 0.851
管道防护措施情况	3	0.697	159.861	3	<0.001	0.814 ~ 0.833
应急响应水平	3	0.676	131.696	3	<0.001	0.777 ~ 0.833
管道管理水平	3	0.560	91.386	3	<0.001	0.526 ~ 0.833

2.3 风险因子初选

风险因子初筛结合相关系数与变异系数考虑。当两因子相关系数较高而存在信息重叠时,保留变异系数较大的因子。高变异系数风险因子具有更离散的样本信息,代表更丰富的信息分布,通常关联极端风险情景,对提升风险预警机制具有重要意义。

由于每个候选风险因子的评分均为 1~5 分,量纲相同,可直接计算相关系数。设 X_{ij} 为第 i 个风险因子的第 j 个样本评分,风险因子 $i = 1, 2, \dots, m$, 样本数据 $j = 1, 2, \dots, n$ 。对于任意两个风险因子 X_i 、 X_k , 相关系数 ρ 计算公式为

$$\rho_{ik} = \frac{\sum_{j=1}^n (X_{ij} - \bar{X}_{ij})(X_{kj} - \bar{X}_{kj})}{\sqrt{\sum_{j=1}^n (X_{ij} - \bar{X}_{ij})^2} \sqrt{\sum_{j=1}^n (X_{kj} - \bar{X}_{kj})^2}} \quad (1)$$

式中 ρ_{ik} 为风险因子 X_i 和 X_k 的相关系数; $\bar{X}_{ij}$ 、 $\bar{X}_{kj}$ 为其样本均值;分子为两因子样本的协方差,分母为两因子样本的标准差乘积。

变异系数衡量风险因子的信息贡献度,计算公式为

$$V_i = \frac{\sqrt{\frac{1}{n-1} \sum_{j=1}^n (X_{ij} - \frac{1}{n} \sum_{j=1}^n X_{ij})^2}}{\frac{1}{n} \sum_{j=1}^n X_{ij}} \quad (2)$$

式中 V_i 为风险因子 X_i 的变异系数,等于其标准差与均值之比。

初步筛选步骤如下。

步骤 1:将候选风险因子数据代入式(1),计算相关系数矩阵,同一维度内因子相关性最强,跨维度因子亦存在较强相关性,需筛选冗余信息。

步骤 2:通过统计学标准与实际数据特征综合确定风险因子筛选的相关系数阈值 P 。现有研究普遍认为皮尔逊相关系数绝对值处于 0.5~0.6 区间表征中强相关水平,本文候选风险因子间相关系数分布于 -0.004~0.613。基于理论模型中强相关判断依据,同时考虑实际数据离散特征,设定相关系数阈值 P 为 0.53。

步骤 3:将候选风险因子数据代入式(2),计算变异系数 V_i ,并排序,结果见表 4。

步骤 4:候选风险因子两两比选。系统随机抽样两项风险因子进行相关系数比对,若二者相关系数绝对值超过预设阈值 P ,则剔除变异系数较低者,保留较高者继续参与后续迭代比选。反之,若二者相关系数绝对值未达阈值标准,则均保留至后续迭代流程。

步骤 5:循环步骤 4 至所有风险因子间的相关系数均低于阈值,终止循环并输出初选风险因子集合,筛选结果见表 4。

2.4 风险因子终选

在初选风险因子体系的基础上,进一步剔除信息冗余贡献度高的因子以提升代表性。双阶皮尔逊关联法第二阶段采用整体相关性法,计算剔除各因子前后体系整体相关度的变化,量化各因子对整体信息冗余的贡献方向与大小。具体而言,整体相关度为初选风险因子相关系数绝对值的平均值。剔除某一因子后,剩余因子相关系数绝对值的平均值构成新的整体相关度。信息冗余贡献指数通过对比剔除前后整体相关度的相对变化,衡量单因子的冗余程度,计算说明如下。

表4 风险因子初选结果
Table 4 Primary selection results of the risk factors

二级维度	风险因子	暂时保留情况	变异系数
暴雨气象特征	降雨量(X_1)	删除	0.275 6
	降雨强度(X_2)	保留	0.248 0
	降雨持续时间(X_3)	删除	0.260 6
	降雨频率(X_4)	保留	0.292 4
	小时最大降雨量(X_5)	保留	0.297 6
	短历时强降雨频次(X_6)	删除	0.268 6
	暴雨发生的季节(X_7)	保留	0.308 8
地形地貌条件	地形坡度(X_8)	删除	0.266 4
	植被覆盖度(X_9)	保留	0.282 7
	水体分布(X_{10})	保留	0.291 0
	沟谷分布(X_{11})	删除	0.267 4
地质条件	土壤类型(X_{12})	保留	0.244 1
	断层分布(X_{13})	删除	0.236 0
	地下水位(X_{14})	删除	0.241 8
	岩性(X_{15})	保留	0.272 4
管道自身情况	管道埋深(X_{16})	删除	0.267 0
	管道壁厚(X_{17})	删除	0.247 2
	管道直径(X_{18})	删除	0.264 3
	管道材质与抗压强度(X_{19})	保留	0.272 9
	管道服役年限(X_{20})	保留	0.291 4
	管道防腐层状况(X_{21})	删除	0.259 7
	管道运行内压(X_{22})	删除	0.253 2
	管道敷设方式(X_{23})	保留	0.284 1
管道周边社会环境	周边交通设施分布(X_{24})	删除	0.274 0
	人口密度(X_{25})	保留	0.313 3
气象监测与预警水平	气象监测站点覆盖密度(X_{26})	保留	0.325 2
	暴雨预警准确率(X_{27})	删除	0.298 4
	预警信息传递效率(X_{28})	保留	0.334 4
管道防护措施情况	排水设施建设情况(X_{29})	删除	0.231 0
	水土保持措施情况(X_{30})	保留	0.231 2
	抗灾加固措施情况(X_{31})	保留	0.266 6
应急响应水平	应急物资储备情况(X_{32})	保留	0.273 8
	应急演练频率与效果(X_{33})	保留	0.252 2
	与周边单位应急联动能力(X_{34})	保留	0.271 7
管道管理水平	巡检频率与覆盖范围(X_{35})	删除	0.289 3
	管理人员培训与专业能力(X_{36})	保留	0.333 7
	信息化管理水平(X_{37})	保留	0.290 8

单个风险因子相关度是风险因子 X_i 与其他 $h - 1$ 个初选因子的相关系数绝对值之和,计算公式为

$$S_i = \sum_{\substack{k=1 \\ k \neq i}}^h |\rho_{ik}| \quad (3)$$

式中 S_i 为风险因子 X_i 的单个风险因子相关度; ρ_{ik} 为 X_i 与配对风险因子 X_k 的相关系数。

整体相关度为所有风险因子单因子相关度的平均值,计算公式为

$$C_{\text{total}} = \frac{1}{h} \sum_{i=1}^h S_i = \frac{1}{h} \sum_{i=1}^h \sum_{\substack{k=1 \\ k \neq i}}^h |\rho_{ik}| \quad (4)$$

式中 C_{total} 为所有风险因子间的整体相关度; h 为初选风险因子个数。

剔除风险因子 X_i 时,剩余 $h - 1$ 个因子的整体相关度计算公式为

$$C_{-p} = \frac{1}{h-1} \sum_{\substack{v=1 \\ v \neq i}}^h \sum_{\substack{s=1 \\ s \neq v, i}}^h |\rho_{vs}| \quad (5)$$

式中 C_{-p} 为剔除风险因子 X_i 后的整体相关度; ρ_{vs} 为 X_v 与配对因子 X_s 的相关系数, X_v 、 X_s 分别为剩余风险因子中的主因子和配对因子。

信息冗余贡献指数表示风险因子对系统冗余度的影响,计算公式为

$$I_p = \frac{C_{-p} - C_{\text{total}}}{C_{\text{total}}} \quad (6)$$

式中 I_p 为风险因子 X_i 的信息冗余贡献指数。

整体相关度与信息冗余贡献指数变化见图 3。当剔除风险因子 X_i 后整体相关度下降 ($C_{-p} < C_{\text{total}}$), 对应 $I_p < 0$, 表明该因子增加整体信息冗余, 且 I_p 越小增加作用越强, 该因子越不重要; 当剔除风险因子 X_i 后整体相关度增大 ($C_{-p} > C_{\text{total}}$), 对应 $I_p > 0$, 表明该因子抑制整体信息冗余, 且 I_p 越大抑制作用越强, 该因子越重要。

图 3 中, C_{total} 代表全部风险因子的整体相关度, C_{-p} 为剔除风险因子 X_i 后的整体相关度, I_p 为该因子的信息冗余贡献指数。图 3 选取的两个点是随机两个因子的信息冗余贡献指数, 仅用于展示其正负属性。 C_{-p} 列的箱形图表示剔除各因子后的整体相关度分布。

对初选风险因子进行信息冗余贡献指数筛选步骤如下。

步骤 1: 构建初选风险因子的相关系数矩阵, 元素 ρ_{ik} 表示风险因子 X_i 、 X_k 的皮尔逊相关系数。对角线元素 ($i = k$ 时) 不参与计算, 因为风险因子与自身的相关系数恒为 1。

步骤 2: 根据式 (3) 和 (4) 计算 21 个初选风险因

子的单因子相关度及整体相关度, 根据式 (5) 计算分别剔除各个风险因子后的整体相关度。

步骤 3: 计算初选风险因子的信息冗余贡献指数, 结果见表 5, 排序见图 4。

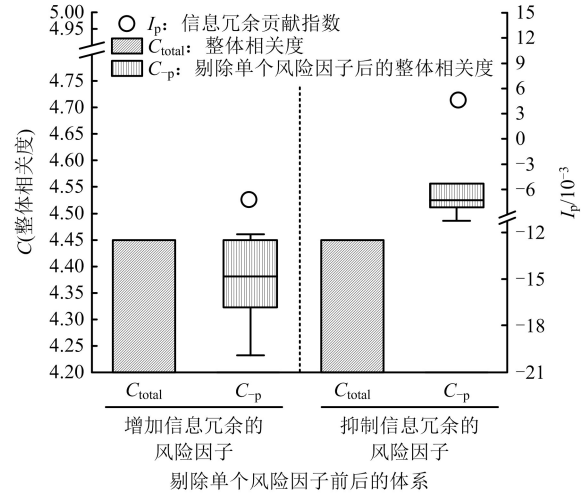


图 3 整体相关度与信息冗余贡献指数的关系

Fig. 3 Relationship between overall correlation and information-redundancy contribution index

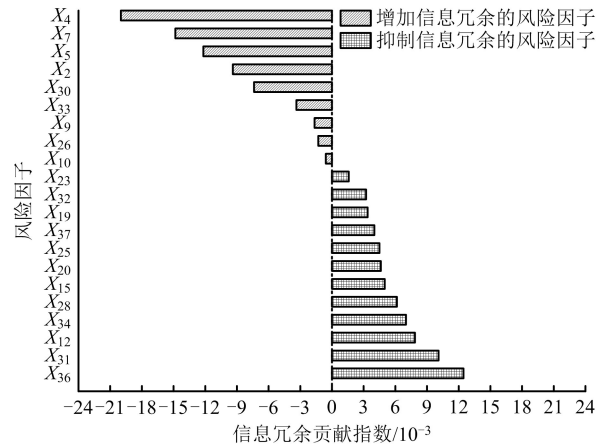


图 4 初选风险因子的信息冗余贡献指数排序

Fig. 4 Ranking of information-redundancy contribution index for primary risk factors

根据主成分分析理论, 当主成分累计方差贡献率超过 85%, 即可认为具备较强解释力。因此, 当累计信息贡献率达到 85% 时, 可判定终选风险因子体系有效保留了初选因子信息, 体系构建合理^[27]。累计信息贡献率 O_i 由终选与初选风险因子变异系数之和的比值计算, 公式为

$$O_i = \frac{U_F}{U_p} \quad (7)$$

式中 U_F 为终选风险因子的变异系数之和; U_p 为

初选风险因子的变异系数之和。

根据信息冗余贡献指数排序,依次选取风险因子,直至累计信息贡献率 O_i 超过 85%,终选结果见表 5。

2.5 稳定性检验

为验证本文方法的稳定性,随机将问卷数据划分为两组进行风险因子筛选。采用 Jaccard 相似系数评估两个风险因子体系的一致性。基于理论标准,当 $0.60 \leq J \leq 0.75$ 时,表明本文方法稳定性较强^[28]。计算公式为

$$J(A, B_w) = \frac{|A \cap B_w|}{|A \cup B_w|} \quad (8)$$

式中 $J(A, B_w)$ 为原始风险因子体系 A 与分组后体系 $B_w (w = 1, 2)$ 的 Jaccard 相似系数。

根据研究,当风险因子超过 15 个时,各因子权重差异较小^[29],因此假设其影响相同以简化计算。

结果显示,第一组筛选出 17 个风险因子,与未分组体系的 Jaccard 相似系数为 71.4%,信息解释强度为 1.06,累计信息贡献率为 85.8%;第二组筛选出 18 个风险因子, Jaccard 系数为 60.9%,信息解释强度为 1.03,累计贡献率达 90.3%。两组结果均满足稳定性判断标准,验证了本文方法的稳健性。

风险因子体系的信息解释强度指数为终选风险因子数据方差和的均值与候选风险因子方差和均值之比,值越大说明信息解释能力越强,计算公式为

$$N = \frac{\frac{1}{t} \sum_{i=1}^t \frac{\sum_{j=1}^n (X_{ij} - \bar{X}_{ij})^2}{n-1}}{\frac{1}{m} \sum_{i=1}^m \frac{\sum_{j=1}^n (X_{ij} - \bar{X}_{ij})^2}{n-1}} \quad (9)$$

式中 N 为风险因子体系的信息解释强度指数; t

表 5 风险因子终选结果

Table 5 Final selection results of risk factors

二级维度	风险因子	最终保留情况	信息冗余贡献指数/ 10^{-3}
暴雨气象特征	降雨强度(X_2)	保留	-9.33
	降雨频率(X_4)	删除	-19.84
	小时最大降雨量(X_5)	保留	-12.17
	暴雨发生的季节(X_7)	删除	-14.68
地形地貌条件	植被覆盖度(X_9)	保留	-1.54
	水体分布(X_{10})	保留	-0.55
地质条件	土壤类型(X_{12})	保留	7.85
	岩性(X_{15})	保留	4.99
管道自身情况	管道材质与抗压强度(X_{19})	保留	3.37
	管道服役年限(X_{20})	保留	4.62
	管道敷设方式(X_{23})	保留	1.58
管道周边社会环境	人口密度(X_{25})	保留	4.50
气象监测与预警水平	气象监测站点覆盖密度(X_{26})	保留	-1.23
	预警信息传递效率(X_{28})	保留	6.14
管道防护措施情况	水土保持措施情况(X_{30})	保留	-7.24
	抗灾加固措施情况(X_{31})	保留	10.10
应急响应水平	应急物资储备情况(X_{32})	保留	3.23
	应急演练频率与效果(X_{33})	保留	-3.31
	与周边单位应急联动能力(X_{34})	保留	7.01
管道管理水平	管理人员培训与专业能力(X_{36})	保留	12.45
	信息化管理水平(X_{37})	保留	4.01

为终选风险因子数, m 为候选风险因子数。

2.6 案例验证

为验证终选风险因子体系的合理性与适用性,选取 2020 年贵州遵义桐梓中石化西南成品油管道柴油泄漏事故(中华人民共和国生态环境部《次生重大突发环境事件调查报告》,2020-10-29)作为典型案例进行对照分析。该事故因持续性强降雨诱发山体滑坡,导致管道变形破裂并发生大规模柴油泄漏,造成跨省重大环境污染。事故的成因与演化过程能够较好地检验终选风险因子体系的覆盖性与解释力。案例与终选风险因子的对应关系见表 6。

综上,暴雨下桐梓管道受灾事故的演化过程几乎涵盖了全部终选风险因子。强降雨与复杂地形地

质条件共同触发灾害,管道特征、防护措施和管理薄弱加剧损害,监测预警缺陷及应急响应不足进一步放大后果,充分验证了终选风险因子体系的合理性与适用性。

3 方法对比

为验证本文方法的信息冗余筛选效果,采用 K -中心点聚类 and K -均值聚类算法对候选风险因子聚类,并从每类中选取变异系数最大的因子构建风险因子体系。通过对比三种方法在因子分布、信息解释强度指数和累计信息贡献率方面的表现,验证本文构建的暴雨天气下长输管道受灾风险因子体系的合理性。

表 6 终选风险因子在案例中的适用情况
Table 6 Application of final risk factors in case studies

二级维度	风险因子	案例印证点
暴雨气象特征	降雨强度(X_2)	持续强降雨长期浸润土体,削弱管基稳定性
	小时最大降雨量(X_5)	短时强降雨产生瞬时冲击,引发管道破坏
地形地貌条件	植被覆盖度(X_9)	滑坡区植被覆盖度不足,削弱坡体稳定性
	水体分布(X_{10})	灾害发生区存在捷阵溪、松坎河等河流,加剧滑坡发生风险,并在灾害发生后扩大污染范围
地质条件	土壤类型(X_{12})	碎屑土层易受雨水浸润软化,降低土壤强度
	岩性(X_{15})	不利岩土结构(软弱夹层)促成滑坡形成
管道自身情况	管道材质与抗压强度(X_{19})	管道为钢质,设计压力为 9.5 MPa,外部载荷超限导致受灾
	管道服役年限(X_{20})	管道服役多年,抗风险能力下降
	管道敷设方式(X_{23})	管道埋设于滑坡体内,敷设方式影响受灾风险
管道周边社会环境	人口密度(X_{25})	处于居民区附近,灾害影响范围扩大,应急处置难度增加
气象监测与预警水平	气象监测站点覆盖密度(X_{26})	当地气象监测不足,未及时发布预警
	预警信息传递效率(X_{28})	政企信息通报延迟,影响应急处置时效
道防护措施情况	水土保持措施情况(X_{30})	水土保持措施不足,增加滑坡风险
	抗灾加固措施情况(X_{31})	未对已知滑坡隐患区采取加固措施
应急响应水平	应急物资储备情况(X_{32})	应急物资准备不足,限制快速处置能力
	应急演练频率与效果(X_{33})	应急演练不足,盲目开挖管道导致泄漏扩大
	与周边单位应急联动能力(X_{34})	各单位间信息通报不及时,跨区域联动不足
管道管理水平	管理人员培训与专业能力(X_{36})	企业对风险判断不足,抢险中多次失误
	信息化管理水平(X_{37})	数据采集与监视控制(Supervisory Control and Data Acquisition,SCADA)系统虽运行,但未能有效判断风险

3.1 筛选结果对比

以 K -中心点聚类法和 K -均值聚类算法为对比方法,将候选风险因子划分为与本文终选数相同的 19 个簇,并选取每簇变异系数最大的因子作为代表,构

建两个包含 19 个因子的风险因子体系。本文方法与两种聚类方法的筛选结果对比见表 7。

分析结果表明, K -中心点聚类算法筛选出的风险因子体系存在维度缺失与分布失衡问题,“管道周

表 7 三种方法筛选结果对比
Table 7 Comparison of screening results from three methods

二级维度	风险因子	双阶皮尔逊 关联法	K -中心点 聚类算法	K -均值 聚类算法
暴雨气象特征	降雨量(X_1)	删除	保留	删除
	降雨强度(X_2)	保留	删除	删除
	降雨持续时间(X_3)	删除	删除	删除
	降雨频率(X_4)	删除	删除	删除
	小时最大降雨量(X_5)	保留	删除	保留
	短历时强降雨频次(X_6)	删除	删除	保留
	暴雨发生的季节(X_7)	删除	删除	保留
地形地貌条件	地形坡度(X_8)	删除	保留	删除
	植被覆盖度(X_9)	保留	保留	删除
	水体分布(X_{10})	保留	删除	保留
	沟谷分布(X_{11})	删除	删除	删除
地质条件	土壤类型(X_{12})	保留	保留	保留
	断层分布(X_{13})	删除	删除	删除
	地下水位(X_{14})	删除	保留	删除
	岩性(X_{15})	保留	删除	保留
管道自身情况	管道埋深(X_{16})	删除	删除	保留
	管道壁厚(X_{17})	删除	保留	删除
	管道直径(X_{18})	删除	保留	删除
	管道材质与抗压强度(X_{19})	保留	保留	删除
	管道服役年限(X_{20})	保留	保留	保留
	管道防腐层状况(X_{21})	删除	保留	删除
	管道运行内压(X_{22})	删除	保留	删除
	管道敷设方式(X_{23})	保留	保留	保留
管道周边社会环境	周边交通设施分布(X_{24})	删除	删除	删除
	人口密度(X_{25})	保留	删除	保留
气象监测与预警水平	气象监测站点覆盖密度(X_{26})	保留	删除	保留
	暴雨预警准确率(X_{27})	删除	保留	保留
	预警信息传递效率(X_{28})	保留	保留	保留
管道防护措施情况	排水设施建设情况(X_{29})	删除	删除	删除
	水土保持措施情况(X_{30})	保留	保留	保留
	抗灾加固措施情况(X_{31})	保留	保留	保留
应急响应水平	应急物资储备情况(X_{32})	保留	保留	保留
	应急演练频率与效果(X_{33})	保留	删除	删除
	与周边单位应急联动能力(X_{34})	保留	删除	保留
管道管理水平	巡检频率与覆盖范围(X_{35})	删除	删除	删除
	管理人员培训与专业能力(X_{36})	保留	保留	保留
	信息化管理水平(X_{37})	保留	保留	保留

边社会环境”维度因子缺失,“管道自身情况”维度占比达 36.8%。 K -均值聚类算法亦分布不均,“地形地貌条件”仅保留“水体分布”一项。相比之下,本文方法在保留各维度基础上实现风险因子分布均衡,有效避免了因子过度集中,构建出更全面、合理的风险因子体系。

3.2 筛选效果对比

3.2.1 信息解释强度指数和累计信息贡献率对比

对比三种方法筛选的风险因子体系,在累计信息贡献率上,本文方法为 89.9%,显著高于 K -中心点聚类算法(51.2%)和 K -均值聚类算法(53.7%),信息保留更充分。在信息解释强度指数方面,三者分别为 1.033、1.007 和 1.057, K -均值聚类算法略高。但 K -均值聚类算法在稳定性检验中,筛选结果与分组前的 Jaccard 相似系数均低于 0.60,稳定性较差,可靠性不足。

3.2.2 KMO 变化和整体冗余度对比

图 5 显示三种方法筛选后风险因子体系的 KMO 值变化。本文方法将 KMO 值由 0.889 降至 0.780,降幅为 12.3%; K -均值聚类算法降至 0.793,降幅为 10.7%; K -中心点聚类算法降至 0.834,降幅仅为 6.2%。结果表明,本文方法在降低信息冗余方面效果最优。通过对比筛选前后风险因子的相关性,可直观反映信息冗余变化。候选风险因子体系的相关系数热力图见图 6(a),双阶皮尔逊关联法、 K -中心点聚类法和 K -均值聚类算法筛选结果分别见图 6(b)、(c)和(d)。分析表明,候选风险因子体系存在明显强相关,同一维度相关性最强,跨维度亦较

高。经双阶皮尔逊关联法筛选后,同维度和跨维度相关性均显著降低;而 K -中心点聚类算法保留了同维度的高相关性, K -均值聚类算法整体相关性仍高于本文方法。综上,双阶皮尔逊关联法在降低信息冗余方面效果更优。

3.2.3 方法适用场景对比

数据分析表明,候选风险因子体系规模较小、信息冗余高、风险因子同质性强,易引入噪声,且终选风险因子目标个数不确定。 K -中心点和 K -均值聚类算法受限于预设簇数,且对噪声敏感,候选因子较少时易导致过度聚集、簇内稀疏或空簇,影响代表性。相比之下,本文方法通过相关系数-变异系数双重准则及信息冗余贡献指数的两阶段筛选,有效剔除信息冗余,构建出信息解释强度和累计贡献率均达标的风险因子体系。

4 讨 论

在研究过程中发现,剔除某一特定风险因子后,体系整体相关性反而上升,表现出明显的信息涌现特征。信息涌现是指系统整体属性无法由其个体特征简单推导^[30],结合因果涌现理论与因子相关性分析,可通过相关性变化衡量系统的涌现强度^[31-32]。本文中多数风险因子间存在中强相关性,具备信息涌现的结构基础。计算结果显示,约 57.14% 的风险因子在被剔除后引发体系整体相关性增强,其信息冗余贡献指数为正,表明该类因子在体系中起“缓冲”作用,抑制了其他因子的相关性,是可能的“共同驱动因子”。以“管理人员培训与专业能力”为例,剔除后可能增强“应急物资储备情况”“应急演练频率与效果”“与周边单位应急联动能力”等因子之间的关联性,原本集中于该因子的冗余信息被重新分配,导致体系整体相关性上升。该现象印证了 Rosas 等^[33]提出的观点:冗余信息可能掩盖系统涌现,剔除后反而增强系统内在联系。因此,信息涌现理论不仅有助于揭示风险因子间的协同机制,也可用于评估信息冗余程度,并可将因子对系统涌现的贡献纳入权重计算,进一步提升风险评估模型的科学性与精准性。本研究仍存在一些待完善之处。在数据层面,样本虽具代表性但未覆盖全国区域,区域间对因子的关注点可能存在差异;数据主要依赖专家经验,未来将引入管道监测与历史气象数据进行多维验证。在方法层面,双阶皮尔逊关联法对阈值设置较为敏感,当因子相关性接近时筛选效果有限。

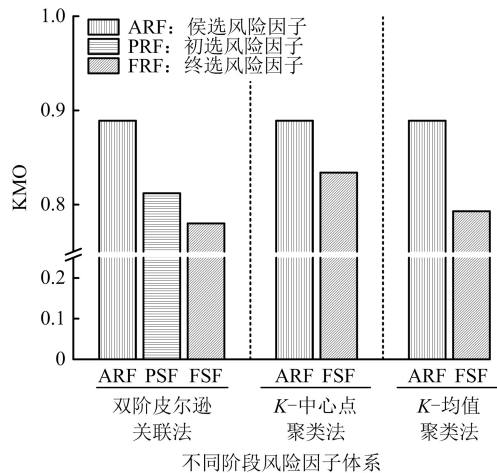


图 5 三种方法筛选前后风险因子体系 KMO 对比

Fig. 5 KMO comparison of the risk factor system before and after screening with three methods

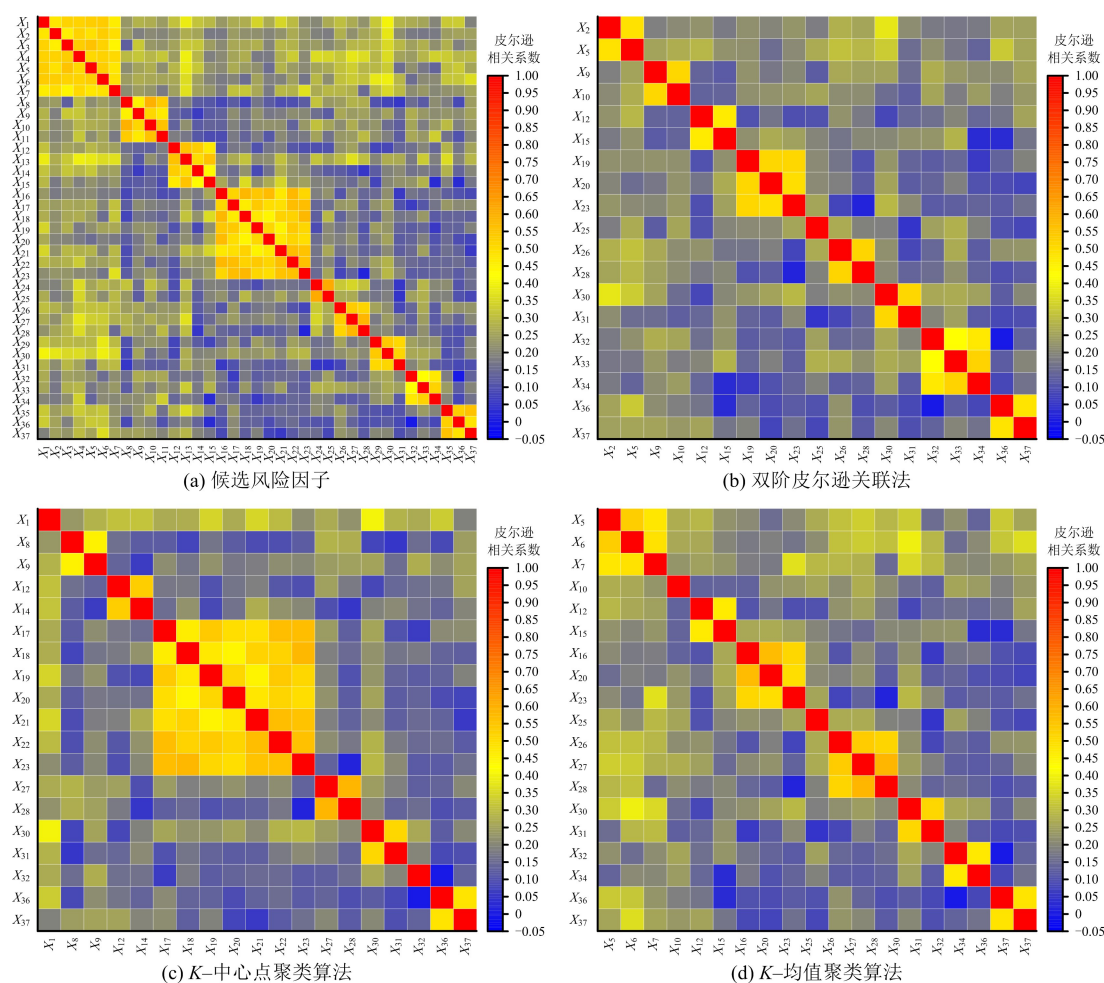


图6 三种方法筛选前后风险因子相关性热力图对比

Fig. 6 Comparison of risk factors correlation heatmaps before and after screening by three methods

在应用层面,目前风险因子体系主要适用于陆地管道,对海底管道等特殊场景尚缺适应性。

5 结论

本文基于局部-整体信息冗余最优角度,以暴雨天气下长输管道为研究对象,构建了涵盖气象、地形、地质、管道自身情况、社会环境等九个维度的风险因子体系,筛选出37个候选风险因子。提出双阶皮尔逊关联法进行信息冗余与信息贡献度双重筛选,并与K-中心点聚类、K-均值聚类算法进行对比,借助信息解释强度指数与累计信息贡献率评估风险因子体系合理性,从整体信息冗余变化与风险因子分布均衡性进行多维验证,主要结论如下。

(1)基于灾害链理论,构建了九个维度的风险因子体系,覆盖暴雨天气下长输受灾全过程,弥补以往在灾前防控、管道运维管理与应急处置等环节的风险因子缺失。

(2)采用双阶皮尔逊关联法,筛选出终选风险因子体系的信息解释强度指数为1.033,累计信息贡献率达89.9%,风险因子信息贡献度高且信息冗余度低,既控制冗余信息又避免有效信息损失,实现最优信息筛选效果。

(3)多维对比分析显示,本文方法在信息解释强度指数、累计信息贡献率、信息冗余控制与风险因子均衡性方面均优于传统聚类方法,验证了其科学性与有效性。实际应用中应结合区域特点动态调整风险因子权重,并融合实时与历史数据以提升评估准确性。

参考文献 (References):

- [1] LORENZO R, VALERIA M C, GABRIELE L. Vulnerability assessment of process pipelines affected by flood events [J]. Reliability Engineering and System Safety, 2022, 219: 108261.

- [2] 高冬辰,伍国军,杨典森. 滑坡区域管道不同穿越模式的力学响应与影响因素分析[J]. 武汉大学学报(工学版), 2024, 57(8): 1077–1087.
GAO D C, WU G J, YANG D S. Mechanical response analysis and influencing factors of different crossing modes of pipeline in landslide area[J]. Engineering Journal of Wuhan University, 2024, 57(8): 1077–1087.
- [3] 丁鸿超,蒋宏业,赵新好,等. 黏性泥石流对山区油气管道的冲击动力响应研究[J]. 防灾减灾工程学报, 2023, 43(2): 277–285.
DING H C, JIANG H Y, ZHAO X H, et al. Study on the impact dynamic response of viscous debris flow on oil and gas pipelines in mountainous areas [J]. Journal of Disaster Prevention and Mitigation Engineering, 2023, 43(2): 277–285.
- [4] 钟威,高剑锋. 油气管道典型地质灾害危险性评价[J]. 油气储运, 2015, 34(9): 934–938.
ZHONG W, GAO J F. Hazard assessment of typical geological disasters along oil and gas pipeline[J]. Oil & Gas Storage and Transportation, 2015, 34(9): 934–938.
- [5] BADIDA P, BALASUBRAMANIAM Y, JAYAPRAKASH J. Risk evaluation of oil and natural gas pipelines due to natural hazards using fuzzy fault tree analysis[J]. Journal of Natural Gas Science and Engineering, 2019, 66: 284–292.
- [6] 李越,刘波. 油气长输管道建设中地质灾害风险管理的研究与应用:以阆中-南充输气管道为例[J]. 灾害学, 2018, 33(1): 152–155, 161.
LI Y, LIU B. Research and application of geological disaster risk management in construction of long-distance oil and gas pipelines: a case study of Langzhong-Nanchong gas pipelines[J]. Journal of Catastrophology, 2018, 33(1): 152–155, 161.
- [7] 晋良海,姜桂莲,易小钰,等. 长距离管道水毁灾害风险的半结构化评价研究[J]. 人民长江, 2021, 52(1): 13–19.
JIN L H, JIANG G L, YI X Y, et al. Research on semi-structured risk evaluation of water damage disaster in long-distance pipelines [J]. Yangtze River, 2021, 52(1): 13–19.
- [8] 冼国栋,刘奎荣,吴森,等. 基于GIS的兰成原油管道地质灾害风险评价[J]. 油气储运, 2019, 38(4): 379–384.
XIAN G D, LIU K R, WU S, et al. GIS based risk assessment on geological disasters along Lanzhou-Chengdu crude oil pipeline [J]. Oil & Gas Storage and Transportation, 2019, 38(4): 379–384.
- [9] HAILE W, SOLOMON T. Failure assessment of oil and gas transmission pipelines using an integrated Bayesian belief network and GIS model[J]. International Journal of Pressure Vessels and Piping, 2023, 205: 104984.
- [10] 王灿星,朱杰勇,喻聪骏,等. 基于皮尔逊Ⅲ型曲线的不同降雨工况下的崩滑地质灾害危险性评价[J]. 地质科技通报, 2025, 44(1): 194–204.
WANG C X, ZHU J Y, YU C J, et al. Risk assessment of landslide geological hazards under different rainfall conditions based on the Pearson Ⅲ curves[J]. Bulletin of Geological Science and Technology, 2025, 44(1): 194–204.
- [11] ZAHID H, GOHAR K, BABAR S, et al. On the utility of parents' historical data to investigate the causes of autism spectrum disorder: a data mining-based framework[J]. Innovation and Research in Biomedical Engineering, 2023, 44: 100780.
- [12] PAN D, HAN Y, JIN Q Q, et al. Study of typical electric two-wheelers pre-crash scenarios using K-medoids clustering methodology based on video recordings in China [J]. Accident Analysis and Prevention, 2021, 160: 106320.
- [13] 陈洪海. 基于病态指数循环分析的评价指标筛选研究[J]. 中国管理科学, 2019, 27(1): 184–193.
CHEN H H. Method of screening evaluation indicators based on circle ill-condition index analysis[J]. Chinese Journal of Management Science, 2019, 27(1): 184–193.
- [14] PENG H C, LONG F H, DING C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226–1238.
- [15] WU X F, HUANG X. Screening of urban environmental vulnerability indicators based on coefficient of variation and anti-image correlation matrix method[J]. Ecological Indicators, 2023, 150: 110196.
- [16] 刘威,董婉琪. 基于AHP-熵权法组合赋权的排水管网风险评估方法研究[J]. 安全与环境学报, 2021, 21(3): 949–956.
LIU W, DONG W Q. On the natural gas pipeline disaster consequence evaluation technology based on the attribute interval recognition model[J]. Journal of Safety and Environment, 2021, 21(3): 949–956.

- [17] 熊俊楠, 孙明远, 孙铭. 基于GIS及耦合协调原理的长输管道山洪泥石流风险性评价[J]. 天然气工业, 2019, 39(3): 116–124.
XIONG J N, SUN M Y, SUN M. Risk assessment on mountain torrents and debris flows along long-distance pipelines based on the GIS and coupling-coordination principle[J]. Natural Gas Industry, 2019, 39(3): 116–124.
- [18] 侯志强, 赵鹏, 曾春雷. 石化港区油品管道安全风险评估技术研究[J]. 安全与环境学报, 2019, 19(2): 371–377.
HOU Z Q, ZHAO P, ZENG C L. Probe into the technology for the risk assessment of oil pipeline in the petrochemical port sphere [J]. Journal of Safety and Environment, 2019, 19(2): 371–377.
- [19] 李卫宁, 金舒婷. 基于灾害链理论的危机信息需求预测研究[J]. 情报科学, 2021, 39(9): 25–34.
LI W N, JIN S T. A study on crisis information demand forecasting based on disaster chain theory [J]. Information Science, 2021, 39(9): 25–34.
- [20] 郭昕曜, 星宇铮, 王远声, 等. 新闻文本挖掘下“暴雨-山洪-地质灾害”的情景演化分析[J]. 安全与环境学报, 2025, 25(4): 1466–1476.
GUO X Y, XING Y Z, WANG Y S, et al. Scenario evolution analysis of “rainstorm-flash flood-geological disasters” based on news text mining [J]. Journal of Safety and Environment, 2025, 25(4): 1466–1476.
- [21] WANG Y H, ZHU S Y, WANG B H, et al. Structural health monitoring of oil and gas pipelines: developments, applications and future directions [J]. Ocean Engineering, 2024, 308: 118293.
- [22] KANG J, MENG X X, SU T, et al. Research on leakage control of river oil and gas pipelines based on accident situation evolution model [J]. Journal of Loss Prevention in the Process Industries, 2025, 96: 105615.
- [23] XU T, PANG H X, ZHAN Z J, et al. Characteristics of water vapor isotopes and moisture sources for short-duration heavy rainfall events in Nanjing, eastern China [J]. Journal of Hydrology, 2023, 622: 129731.
- [24] 卢珊, 胡泽勇, 王百朋, 等. 近56年中国极端降水事件的时空变化格局[J]. 高原气象, 2020, 39(4): 683–693.
LU S, HU Z Y, WANG B P, et al. Spatio-temporal patterns of extreme precipitation events over China in recent 56 years [J]. Plateau Meteorology, 2020, 39(4): 683–693.
- [25] ZHANG X X, SONG J X, WANG Y R, et al. Threshold effects of vegetation coverage on runoff and soil loss in the Loess Plateau of China: a meta-analysis [J]. Geoderma, 2022, 412: 115720.
- [26] 伍颖, 查四喜, 金鹏威, 等. 泥石流作用下埋地管道的可靠度分析[J]. 自然灾害学报, 2019, 28(2): 199–206.
WU Y, ZHA S X, JIN P W, et al. Reliability analysis of buried gas pipeline subject to debris flow [J]. Journal of Natural Disasters, 2019, 28(2): 199–206.
- [27] 陈洪海. 基于反映象相关矩阵的评价指标筛选方法研究[J]. 中国管理科学, 2022, 30(11): 149–158.
CHEN H H. Research on the screening method of evaluation indicators based on the anti-image correlation matrix [J]. Chinese Journal of Management Science, 2022, 30(11): 149–158.
- [28] TANG M, KAYMAZ Y, LOGEMAN B L, et al. Evaluating single-cell cluster stability using the Jaccard similarity index [J]. Bioinformatics, 2021, 37(15): 2212–2214.
- [29] 刘婷, 鲁梓萌, 刘翕, 等. 基于灾害链的突发公共事件网络舆情风险评估研究[J]. 情报科学, 2025, 43(1): 69–78.
LIU T, LU Z M, LIU X, et al. Network public opinion risk assessment based on disaster chain theory in public emergencies [J]. Information Science, 2025, 43(1): 69–78.
- [30] YANG M Z, WANG Z P, LIU K W, et al. Finding emergence in data by maximizing effective information [J]. National Science Review, 2025, 12(1): 279.
- [31] ROSAS F E, MEDIANO P A M, GASTPAR M, et al. Quantifying high-order interdependencies via multivariate extensions of the mutual information [J]. Physical Review E, 2019, 100(3): 032305.
- [32] ERIK P H, LARISSA A, GIULIO T. Quantifying causal emergence shows that macro can beat micro [J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(49): 19790–19795.
- [33] ROSAS F E, MEDIANO P A M, JENSEN H J, et al. Reconciling emergences: an information-theoretic approach to identify causal emergence in multivariate data [J]. PLoS Computational Biology, 2020, 16(12): e1008289.

Two-stage Pearson correlation screening method for identifying risk factors in long-distance pipelines during rainstorm disasters

HUANG Xing¹, ZHOU Hongjiao²

(1 School of Economics and Management, Southwest Petroleum University, Chengdu 610500, China)

(2 School of Civil Engineering and Geomatics, Southwest Petroleum University, Chengdu 610500, China)

Abstract: This paper proposes a two-stage Pearson correlation method to optimize the risk factor system for long-distance pipelines during heavy rainfall. The objective is to eliminate both local and overall redundancy among candidate factors, resulting in a concise, independent, and explanatory system. Initially, candidate risk factors were systematically identified by integrating disaster chain theory, academic literature, pipeline safety standards, and accident investigation reports. In the first stage, a dual criterion that combines the Pearson correlation coefficient with the coefficient of variation was used to eliminate local information redundancy among the candidate risk factors. When the correlation coefficient between any two factors exceeded the predefined threshold, their coefficients of variation were compared. The factor with the lower coefficient of variation was removed, while the other was retained for further comparison. This iterative process continued until the correlation coefficients among all pairs of risk factors fell below the threshold, resulting in the establishment of the initial risk factor system. In the second stage, the overall correlation after the removal of each factor and the information redundancy contribution index for each factor were calculated. Based on the cumulative information contribution requirement, factors that contributed significantly to overall redundancy were eliminated, leading to the construction of the final risk factor system. The effectiveness of the proposed method was validated through both case evidence and algorithmic comparisons. A typical disaster case of pipeline failure during heavy rainfall was employed to confirm the rationality and practical applicability of the selected factors. Additionally, comparisons with K -medoids and K -means clustering demonstrated that the refined factor system achieved an interpretation intensity index of 1.033 and a cumulative information contribution rate of 89.9%, significantly higher than the results for K -medoids (1.007, 51.2%) and K -means (1.057, 53.7%). These findings indicate that the proposed method effectively eliminates redundancy while preserving critical information. This approach provides a systematic and objective tool to support the prediction and early warning of long-distance pipeline disasters under extreme rainfall conditions.

Key words: safety systematology; rainstorm disaster; long-distance pipeline; risk factors screening; two-stage Pearson correlation method