Open Journal of Natural Science 自然科学, 2025, 13(2), 305-319
Published Online March 2025 in Hans. https://www.hanspub.org/journal/ojns
https://doi.org/10.12677/ojns.2025.132032

Hans汉斯

地理大数据驱动下的中国地理系统多要素时空特征与灾害预测研究

覃　浩*，高　飞，武新浩

西京学院计算机学院，陕西　西安

收稿日期：2024年12月19日；录用日期：2025年3月5日；发布日期：2025年3月12日

摘　要

随着大数据和人工智能的迅猛发展，地理系统问题在地球科学研究中占据着至关重要的地位。它既蕴含着山川湖海的壮美景观以及气候的动态变迁等自然地理现象，又对人口的分布规律、经济活动的开展以及文化的传承发展等人文地理要素产生着深刻影响，然而，存在着诸多的问题。本文利用地理大数据和数学模型，采用随机森林模型和logic回归检验地形对极端天气的影响，随机森林模型进行复杂的非线性关系刻画不同地理因素与极端天气的关联。采用AHP层次分析法进行土地利用变化特征与结构模型推理和构建，通过分析1990~2020年间中国降水量和土地利用/土地覆被的时空演化特征，探讨地形－气候交互作用对极端天气形成的影响，预测2025~2035年间暴雨灾害脆弱地区，并描述中国土地利用变化的特征与结构。

关键词

地理大数据，Logistic回归，随机森林，层次分析法

Research on the Spatio-Temporal Features of Multiple Elements within China's Geographic System and Disaster Prediction Driven by Geographic Big Data

Hao Qin*, Fei Gao, Xinhao Wu

School of Computer Science, Xijing University, Xi'an Shaanxi

Received: Dec. 19th, 2024; accepted: Mar. 5th, 2025; published: Mar. 12th, 2025

*通讯作者。

文章引用: 覃浩, 高飞, 武新浩. 地理大数据驱动下的中国地理系统多要素时空特征与灾害预测研究[J]. 自然科学, 2025, 13(2): 305-319. DOI: 10.12677/ojns.2025.132032

Abstract

With the rapid development of big data and artificial intelligence, the issues of geographical systems occupy a crucial position in earth science research. It encompasses not only natural geographical phenomena such as the magnificent landscapes of mountains, rivers, lakes and seas and the dynamic changes of climate but also has a profound impact on human geographical elements such as the distribution patterns of population, the conduction of economic activities and the inheritance and development of culture. However, there are numerous problems. In this paper, by utilizing geographical big data and mathematical models, the random forest model and logistic regression are employed to examine the impact of terrain on extreme weather. The random forest model depicts the complex nonlinear relationships between different geographical factors and extreme weather. The Analytic Hierarchy Process (AHP) is adopted to conduct reasoning and construction of the model of the characteristics and structure of land use change. Through analyzing the spatio-temporal evolution characteristics of precipitation and land use/land cover in China from 1990 to 2020, the influence of terrain-climate interaction on the formation of extreme weather is explored, the vulnerable areas of rainstorm disasters from 2025 to 2035 are predicted, and the characteristics and structure of land use change in China are described.

Keywords

Geographic Big Data, Logistic Regression, Random Forest, AHP

1. 引言

地理系统，作为地球上自然与人文要素相互交织、复杂互动的庞大体系，一直是地球科学研究的核心领域。它不仅承载着山川湖海、气候变迁等自然地理现象，还深刻影响着人口分布、经济活动、文化传承等人文地理要素，共同塑造了地球表面的多样性与复杂性。在这一背景下，如何准确、全面地理解和表达地理系统的主导特征，成为了地理学家们长期以来的追求。在过去，受限于技术手段和数据获取能力的限制，地理学家们主要依靠宏观结构和定性分析的方法对地理系统进行研究。正是在这样的背景之下，自然地理分析问题变得尤为重要。面对如此庞大的数据量，传统的数据处理和分析方法已经显得力不从心。如何有效整合和利用这些大数据资源，挖掘出其中蕴含的深层次信息，成为当前地球科学研究面临的重要挑战。

本论文采用数学模型，以解决多要素时空特征与灾害预测，建立灾害能力预测模型、关联性分析模型以及土地利用变化模型。其主要任务包括模型分析、数据清洗、数据处理、特征选择、特征组合、特征提取、特征相关性分析、模型建立和模型比较与测试。在各个任务中，特征选择和模型建立的以自然地理条件和人文地理数据为关键目标。本研究利用丰富的地理数据资源，综合运用多种数据分析与建模技术，力求揭示中国地理系统内在规律与特征。

2. 研究方法

本研究通过整合地理大数据和数学模型，系统地剖析了降水量、土地利用/土地覆被类型的时空演化

特征，深入挖掘了地形–气候相互作用对极端天气形成的影响机制。研究选取中国境内的数据，运用描述性统计方法，分析了降水量和土地覆被类型在时间和空间上的演变关系。在时间维度上，计算了年总降水量、月均降水量和植被总覆盖率等基本统计量，以反映数据的集中趋势和离散程度。空间维度上，利用空间变异系数和皮尔逊相关系数等指标，量化了降水量和土地覆被的空间不均匀性，揭示了其空间分布特征。

研究进一步探讨了暴雨等极端天气事件对人类生产生活的影响。采用近邻匹配平均处理筛选和整合数据，组成新的数据集，并进行数据降维，以反映区域性的地理与气候特征。利用随机森林模型和逻辑回归检验地形高程、降水、气温、经纬度对极端天气的影响，刻画了不同地理因素与极端天气的复杂非线性关系。结果显示，降雨的时空变异性和不可控性最强，土地利用具有一定可控性，而地形最为稳定。通过随机森林模型和逻辑回归，确定了暴雨成灾的临界条件，预测了未来十年暴雨灾害的脆弱地区。

在中国尺度上，研究描述了自然地理特征(如“三级阶梯”、800 mm 等降水量线、秦岭–淮河一线)和人文地理特征(如胡焕庸线)。针对土地利用/土地覆被变化，选取了自然地理特征(如耕地、森林、草地、灌木、湿地、纬度)和人文地理特征(如人口密度、GDP)作为主要参数，进行了权重分析，构建了专家评估矩阵。通过降维处理和 AHP 层次分析法，计算了每个区域的总得分，实现了区域间的比较和分析。研究在各问题分析过程中均辅以相应的检验、评估及可视化手段，确保了研究的科学性和有效性。

3. 研究现状

3.1. 地理大数据在地理系统现状

地理大数据的出现为地理系统研究提供了新的视角和方法。Han 和 Miao (2022)开发了基于观测数据的中国大陆逐日降水数据集，为研究降水的时空分布提供了高分辨率的数据支持。余振等(2022)构建了 1900~2019 年中国土地利用和覆盖变化数据集，为分析土地利用变化提供了丰富的数据资源。这些数据集的开发和应用，显著提高了地理系统研究的精度和深度。此外，汤国安(2019)开发了中国数字高程图(1KM)，为地形分析提供了高分辨率的数据支持。Fang 等(2021)开发了中国近地面气温数据集，并利用机器学习方法进行了气温预测，进一步验证了地理大数据在气候研究中的应用价值。这些研究展示了地理大数据在不同地理要素分析中的广泛应用，为综合研究地理系统提供了坚实的数据基础。

3.2. 机器学习在地理系统现状

机器学习方法在地理系统研究中的应用日益广泛。随机森林(Random Forest)模型因其在处理非线性关系和高维数据方面的优势，被广泛应用于极端天气预测和土地利用变化分析。Yu 等(2022)利用随机森林模型分析了森林扩张对中国土地碳汇的影响，展示了机器学习在生态研究中的潜力。Fang 等(2021)利用随机森林模型对中国近地面气温进行了预测，进一步验证了随机森林在气候数据处理中的有效性。逻辑回归(Logistic Regression)模型在处理二分类问题时表现出色，广泛应用于医学、金融和气象等领域。王灿和王嘉琛(2022)利用逻辑回归模型对中国历史人口空间分布进行了预测，展示了逻辑回归在人口地理研究中的应用。徐新良(2017)利用逻辑回归模型对中国 GDP 空间分布进行了分析，进一步验证了逻辑回归在经济地理研究中的应用价值。

时间序列分析方法在地理系统研究中也被广泛应用。王灿和王嘉琛(2022)利用 Prophet 模型对中国历史人口空间分布进行了预测，展示了时间序列分析在人口地理研究中的应用。徐新良(2017)利用 Prophet 模型对中国 GDP 空间分布进行了分析，进一步验证了时间序列分析在经济地理研究中的应用价值。此外，Liu 等(2005)提出了基于遥感数据的 1 公里网格 GDP 空间化方法，为经济地理研究提供了新的技术手段。黄莹等(2009)基于绿洲土地利用的区域 GDP 公里格网化研究，展示了时间序列分析在区域经济研究中的

应用。Yi 等(2006)提出了基于 GIS 的 GDP 数据像素化方法，进一步验证了时间序列分析在地理数据处理中的应用价值。

4. 模型假设与符号说明

4.1. 模型基本假设

(1) 假设未来降雨的时空分布趋势可根据历史数据进行预测，全球气候变化的影响在模型中已被考虑；
(2) 假设土地利用/土地覆被的变化趋势延续过去的发展规律，除非有重大政策调整，否则不会发生突变；
(3) 假设社会经济因素(如人口密度、GDP)在预测期内保持相对稳定，对模型结果的影响可忽略不计；
(4) 假设模型中的参数和权重在预测期内保持不变，不受外界随机因素的干扰；

4.2. 符号说明

本文所用符号如表 1 所示。

Table 1. Symbol explanation
表 1. 符号说明

符号	含义
CV_{xy}	经纬度(x, y)的空间变异系数
$L(x,y)$	年均降水统计量
$\rho(x,y)$	经纬度(x, y)的皮尔逊相关系数
p	隐状态
$Gini$	$Gini$ 系数
δ	人均绿地覆盖率
γ	绿地率
CI	一致性指标
CR	一致性比率

5. 详细方法

5.1. 时空演化特征

结合大数据技术对地理系统进行综合，为后续的预测提供依据，深入研究全球气候变化背景下中国地理环境的演变。选取中国大陆 0.25˚逐日降水数据集(数据集 3) [1]和中国 0.5˚土地利用和覆盖变化数据集(数据集 4) [2] [3]。本论文采用以下步骤建立描述性统计模型：降水量空间统计量模型：通过加载 1990~2020 年降水量数据集直观展示以经纬度为网格的 3D 年均降水量展示图，接着通过空间变异系数得出降水量与空间有着较强的关联性，并且通过建立可视化经纬度与降水量的折线图表和计算皮尔逊相关系数得到降水量随着经度和纬度的变化有着整体的线性关系，得出我国东南方区域降水较多，西南方以及北方地区降水量相对较少。降水量时间模型建立分析：通过加载 1990~2020 年降水量数据集，经过数据的整合处理得到了我国总降水量的年均变化图，进一步的通过数据集得到了我国在各个月份的月均降水量，得出我国大多数地区在夏季降水相对较多，在冬季降水相对较少。土地使用类型时空模型建立：

通过加载 1990~2019 年土地类型数据集，计算我国各个土地类型空间变异系数随时间的变化，以及可视化我国 1990 年和 2019 年的各种土地类型的分布图。可知我国土地类型随时间变化并不非常敏感，但土地类型的分布不同地区差异较大。

模型建立

(1) 降水量空间统计量模型

计算 1990 年~2020 年，基于空间位置的年均降水统计量公式如下：

$$L(x,y)=\frac{1}{n}\sum_{t=1990}^{2020}P(x,y,t) \tag{1}$$

其中 n 为年份总数为 31 年，x 为经度，y 为纬度，t 为变量从 1990 年至 2020 年，P 函数为对应经纬度的对应年份的降水量，最终计算出某个经纬度下的年均降水量。通过迭代经纬度得各个经纬度下的年均降水量如图 1 所示：

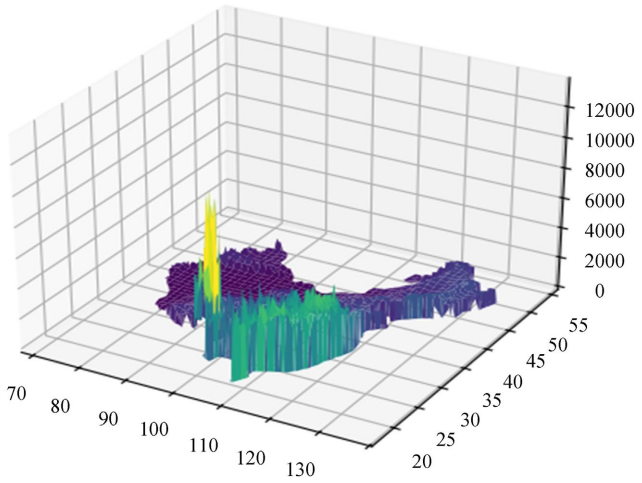

Figure 1. Spatial distribution map of annual average precipitation in spatial location
图 1. 空间位置年均降水量空间分布图

空间变异系数是用来衡量地理空间数据中某个变量的空间分布变化程度的指标。它通常用于描述一个区域内的某种属性(如降水量、温度、土壤含量等)在空间上的离散程度，揭示该属性在不同地点之间的差异性。本论文使用空间变异系数衡量降水量和土地利用类型的空间变化情况。

$$CV_{xy}=\frac{\sigma_{xy}}{\bar{P}_{xy}}\times 100\% \tag{2}$$

$$\sigma_{xy}=\sqrt{\frac{1}{m\times n}\sum_{x=1}^{m}\sum_{y=1}^{n}\left(P(x,y)-\bar{P}_{xy}\right)^2} \tag{3}$$

$$\bar{P}_{xy}=\frac{1}{m\times n}\sum_{x=1}^{m}\sum_{y=1}^{n}P(x,y) \tag{4}$$

式中 $P(x,y)$ 是在经纬度 (x,y) 位置的降水量，m 和 n 分别是经度和纬度的数量。CV 大(通常超过 50%)：表示降水量在不同空间点之间的变化非常显著。不同地区的降水量差异较大。CV 小(通常低于 20%)：表示降水量在不同空间点之间的变化较小，空间分布相对均匀。中间值：表示降水量在空间上有一定程度的变化，但不至于过大。

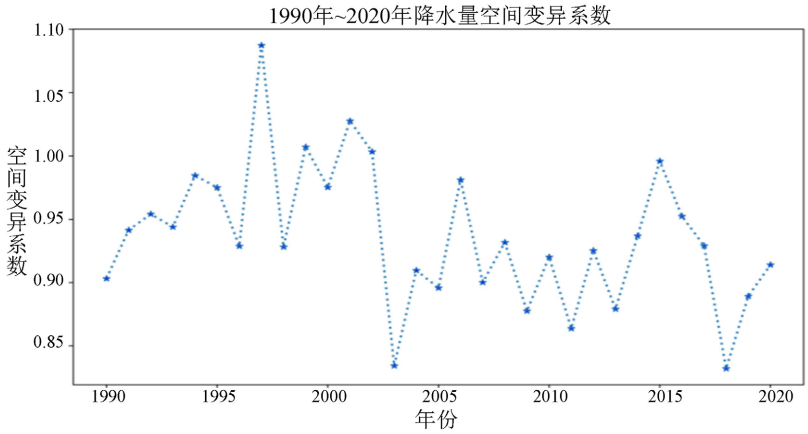

Figure 2. Spatial coefficient of variation of precipitation
图 2. 降水量空间变异系数图

通过图 2 分析空间变异系数可知降水量在不同的空间上有着不同的情况。本文根据此情况将空间分为两个方向以经度方向和以纬度方向，通过经度方向的变化判断此经度上的天均降水量；通过纬度方向变化判断此纬度上的平均降水量，判断经纬度分别对降水量是否成线性关系，本文使用皮尔逊相关系数(Pearson Correlation Coefficient)它是一种用于衡量两个变量之间线性相关程度的统计指标，通常记作 r。它取值范围在−1 到 1 之间，表示两个变量的相关性强弱和方向。$r=1$ 完全正相关，表示两个变量之间存在完美的正线性关系。$r=-1$：完全负相关，表示两个变量之间存在完美的负线性关系。$r=0$：无相关，表示两个变量之间没有线性关系，公式如下：

$$\rho_{X,Y}=corr(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y} \tag{5}$$

式中 X 代表经度/纬度，Y 代表降水量。

Table 2. Pearson coefficients of latitude, longitude and precipitation
表 2. 经纬度与降水量的皮尔逊系数

皮尔逊相关系数 r	经度	纬度
降水量	0.3508522542562407	−0.44367048925004265

通过计算出的相关系数如表 2 可判断在整体情况下随着经度增加时降水量增加，在随着纬度增加时降水量增加，这和图 1 展示的直观结论相吻合。

(2) 降水量时间统计量模型

t 为某一年的年份，在经度 x，纬度 y 的降水量。m 是经度的数量，n 是纬度的数量。Z 的值为某年的总降水量。

$$Z_{total\_year}\left(t_{i\_year}\right)=\sum_{x=1}^{m}\sum_{y=1}^{n}P\left(t_{i\_year},x,y\right) \tag{6}$$

t 为某一月份，在经度 x，纬度 y 的降水量。m 是经度的数量，n 是纬度的数量。H 的值为月均降水量。

$$H_{total\_month\_mean}\left(t_{i\_month}\right)=\sum_{x=1}^{m}\sum_{y=1}^{n}P\left(t_{i\_month},x,y\right)\Big/XY \tag{7}$$

通过上述公式得可视化出降水量随着年份的变化和随着月份的变化走势如图 3 所示。

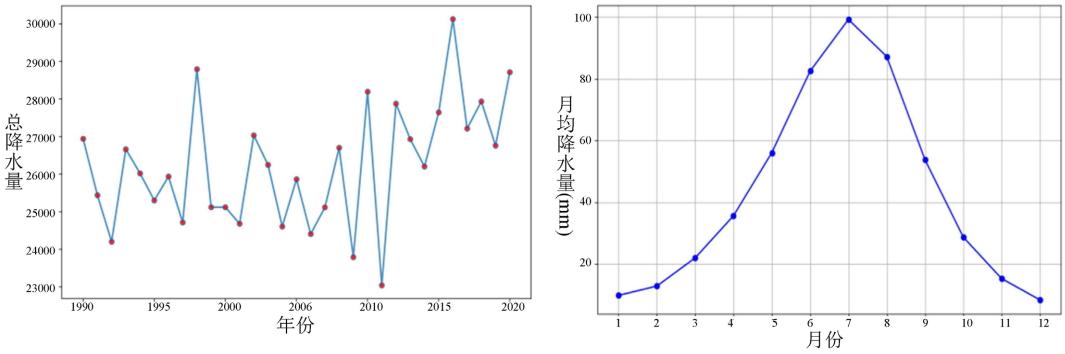

Figure 3. Trend chart of annual total precipitation/monthly average precipitation
图 3. 年份总降水量/月均降水量走势图

通过这两个统计量可以得到不同年份之间降水量波动较大，但在某些月份我国在春夏季整体降水较多，秋冬季降水较少。

5.2. 暴雨成灾特征

5.2.1. 极端天气模型定义

本文首先进行极端类型定义。暴雨模型定义(见表 3)：暴雨分为三种类型：中等暴雨(24 小时降水量 > 50 mm)、强暴雨(24 小时降水量 > 100 mm)、特大暴雨(24 小时降水量 > 200 mm)。热浪模型定义(见表 4)：热浪通常指连续 3 天以上的日最高温度超过当地历史日最高温度的 90%。若日最高温度连续超过 35℃，则可视为较强的热浪。寒潮模型定义(见表 5)：寒潮包括两种情况：急剧降温(24 小时内气温骤降超过 8℃，或 48 小时内气温下降超过 10℃)，以及低温(最低温度连续 2~3 天低于某个极端低温值，如连续 3 天低于 0℃，或低于当地历史最低气温的 10%)。

Table 3. Definition of rainstorm model
表 3. 暴雨模型定义

类型	中等暴雨	强暴雨	特大暴雨
暴雨	24 小时降水量 > 50 mm	24 小时降水量 > 100 mm	24 小时降水量 > 200 mm

Table 4. Definition of heat wave model
表 4. 热浪模型定义

类型	日最高温度	绝对高温
热浪	连续 3 天以上的日最高温度超过当地历史日最高温度 90℃	日最高温度连续超过 35℃，视为较强的热浪。

Table 5. Definition of cold wave model
表 5. 寒潮模型定义

类型	急剧降温	低温
寒潮	24 小时内气温骤降超过 8℃，或 48 小时内气温下降超过 10℃。	最低温度连续 2~3 天低于某个极端低温值。例如，最低温度连续 3 天低于 0℃，或当地历史最低气温 10%以下。

5.2.2. 随机森林模型

随机森林(Random Forest, RF)是一种基于 Bootstrap 随机重采样和随机特征选择的集成学习方法。通

过从原始数据中多次随机抽样构建多棵决策树(Decision Tree)，每棵树在分裂节点时随机选择一部分特征。最终，随机森林通过多个决策树的集成(通常为投票机制)来得到最终的分类结果。作为近年来快速发展的机器学习技术，随机森林在分类、回归、特征选择以及异常检测等任务中得到了广泛应用，并因其灵活性和强大的处理能力备受青睐。

随机森林是多棵决策树的集成，决策树结构图 4 所示。

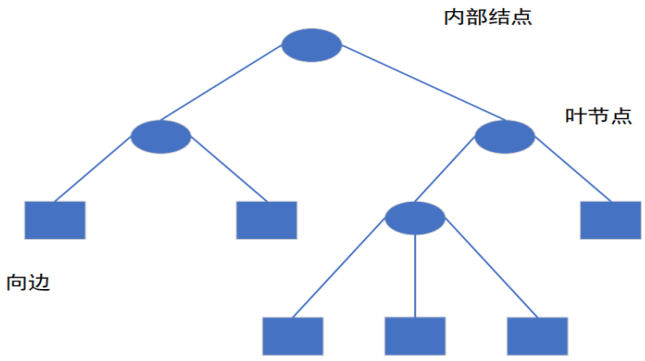

Figure 4. Decision tree structure diagram
图 4. 决策树结构图

众多决策树构成了随机森林，每棵决策树都会有一个投票结果，最终投票结果最多的类别，就是最终的模型预测结果。

随机森林的分类模型具体公式如下：

$$h(x)=\frac{1}{N}\sum_{i=1}^{N}T_i(X) \tag{8}$$

其中： $h(x)$ *是输入* x 的最终预测结果(即是否发生极端天气)；$T_i(X)$ 是第 i 棵决策树的预测输出；

决策树的构建：

1) Bootstrap 抽样：从原始数据集中随机抽取子集作为决策树的训练集。

2) 特征选择：在每个节点进行分裂时，从所有特征中随机选择一个子集，选择子集中分裂效果最好的特征进行分裂。

树的生成过程中，使用信息增益、基尼系数或熵来衡量分裂节点的优劣。

基尼指数公式：对于分类问题，常用基尼指数(Gini Impurity)来衡量不纯度：

$$Gini=1-\sum_{K=1}^{K}P_K^2 \tag{9}$$

5.2.3. 模型建立

(1) 输入变量：

地形变量 X_n：['高程', '降水', '气温', '耕地', '森林', '草地', '灌木', '湿地', '经度', '纬度']

(2) 模型输出：

输出变量 Y：二分类变量，表示是否发生极端天气(1 表示发生，0 表示未发生)。

(3) 建模过程：

使用训练数据 (X_1, X_2, Y) 训练多棵决策树，训练集选择总体的 70%，测试集选择总体的 30%,每棵树使用 Bootstrap 方法抽取训练集样本，并从随机选择的特征子集中选择最优分裂点；

每棵决策树通过递归分裂构建，直至达到停止条件(如最大深度或叶子节点数量)。

(4) 投票机制：

最终通过多数投票的方式决定预测类别。具体公式为：

$$P\left(Y=1\middle|X\right)=\frac{1}{N}\sum_{i=1}^{N}T_i\left(X\right) \tag{10}$$

即所有决策树的输出结果 $T_i\left(X\right)$ 的平均值代表最终的概率，超过一定阈值则分类为 1 (极端天气发生)，否则分类为 0。

模型求解结果如图 5 所示。

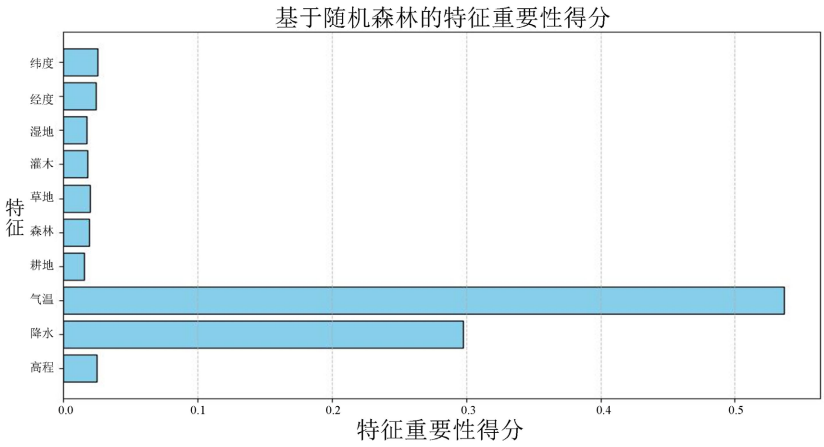

Figure 5. Feature importance score based on random forest
图 5. 基于随机森林的特征重要性得分图

在研究极端天气暴雨成灾形成机制的过程中，随机森林特征重要性得分提供了极具价值的线索，深入探究地形、气候、土地利用等不同因素对极端天气的影响机制及其相互作用。

通过图 5 随机森林算法得出的特征重要性得分，直观地观察到各个因素在极端天气预测模型中的重要程度。在得分较高的因素中，如气候因素，像降水、气温、风速等，占据着重要的地位。降水作为直接与极端天气相关的气候要素，其强度、持续时间和频率的变化直接决定了暴雨、洪涝等灾害的发生。气温的异常波动会导致大气环流的改变，进而影响极端天气的形成。

土地利用类型的变化同样不容忽视。森林、草地、耕地、城市用地等不同土地利用类型，通过改变地表的物理性质和生态过程，对极端天气产生影响。进一步深入分析，我们可以发现地形、气候、土地利用之间存在着复杂的相互作用。地形会影响气候的分布，综上，随机森林特征重要性得分，展现了地形、气候、土地利用等因素对极端天气暴雨成灾的相互作用，后文极端天气的预测、预警提供了理论基础。

5.3. 暴雨成灾预测

5.3.1. 逻辑回归模型

对暴雨数据提取处理后的数据，极端天气仅表示是否出现暴雨。之后使用逻辑回归模型进行拟合。通过拟合，发现 Logit > 0.79 判断为暴雨成灾。

5.3.2. Prophet 时间序列模型

Prophet 是由 Facebook 开发的一种时间序列预测模型，旨在处理具有季节性和趋势性的数据。它通过分解时间序列为趋势、季节性和假期影响三部分来进行建模，能够灵活应对缺失数据和异常值。用户只需提供时间戳和相应的观测值，Prophet 会自动识别数据中的模式并进行预测。模型采用了加法或乘法

的方式来组合这些成分，允许用户对季节性进行定制，并支持假期效应的灵活建模，适用于各种行业的业务需求。使用 Prophet 的过程包括数据准备、模型拟合和未来数据的预测，且其易用性使其受到数据科学家和分析师的广泛欢迎。

模型整体由三部分组成：growth (增长趋势)、seasonality (季节趋势)、holidays (节假日对预测值的影响)：

$$y(t)=g(t)+s(t)+h(t)+\in_{(t)} \tag{11}$$

其中：

$g(t)$ 表示趋势项，它表示时间序列在非周期上面的变化趋势；

$s(t)$ 表示周期项，或者称为季节项，一般来说是以周或者年为单位；

$h(t)$ 表示节假日项，表示时间序列中那些潜在的具有非固定周期的节假日对预测值造成的影响；

即误差项或者称为剩余项，表示模型未预测到的波动，服从高斯分布；

Prophet 算法就是通过拟合这几项，然后最后把它们累加起来就得到了时间序列的预测值。

趋势项：

$$g(t)=\frac{C(t)}{1+\exp\left(-k+a(t)^{T}\delta\right)\cdot\left(t-m+a(t)^{T}\gamma\right)} \tag{12}$$

$$a(t)=\left(a_1(t),\cdots,a_s(t)\right)^{T},\delta=\left(\delta_1,\cdots,\delta_s\right)^{T},\gamma=\left(\gamma_1,\cdots\gamma_s\right)^{T} \tag{13}$$

趋势项有两个重要的函数，一个是基于逻辑回归函数的(非线性增长)，另一个是基于分段线性函数的(线性增长)，$C(t)$ 表示承载量：它是一个随时间变化的函数，限定了所能增长的最大值。

季节性趋势：

$$s(t)=\sum_{n=1}^{N}\left(a_n\cos\left(\frac{2\pi nt}{p}\right)+b_n\sin\left(\frac{2\pi nt}{p}\right)\right) \tag{14}$$

由于时间序列中有可能包含多种天，周，月，年等周期类型的季节性趋势，因此，傅里叶级数可以用来近似表达这个周期属性。

使用傅立叶级数来模拟时间序列的周期性：假设 P 表示时间序列的周期，P = 365.25 表示以年为周期，$P=7$ 表示以周为周期。

它的傅立叶级数的形式都是：N 表示希望在模型中使用的这种周期的个数，较大的 N 值可以拟合出更复杂的季节性函数，然而也会带来更多的过拟合问题。

按照经验值，对于以年为周期的序列(P = 365.25)而言，N = 10；

$$X(t)=\left[\cos\left(\frac{2\pi(1)t}{365.25}\right),\cdots,\sin\left(\frac{2\pi(10)t}{365.25}\right)\right] \tag{15}$$

对于以周为周期的序列(P = 7)而言，N = 3

$$X(t)=\left[\cos\left(\frac{2\pi(1)t}{7}\right),\cdots,\sin(7)\right] \tag{16}$$

因此时间序列的季节项就是：

$$s(t)=X(t)\beta \tag{17}$$

在代码里面，seasonality _ mode 也对应着两种模式，分别是加法和乘法，默认是加法的形式。

本文使用 Prophet 模型进行预测 2025~2035 年的降水量如图 6 所示：

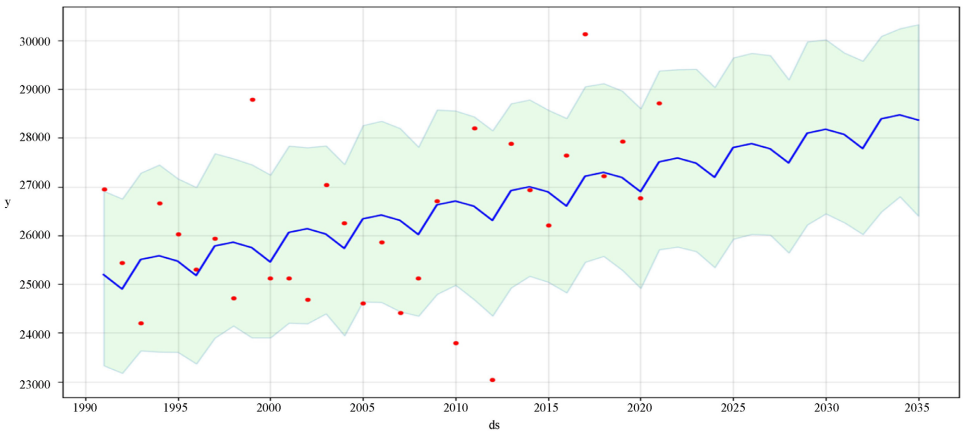

Figure 6. Precipitation map from 2025 to 2035
图 6. 2025~2035 年的降水量图

通过 Prophet 时间序列模型预测后的数据使用逻辑回归模型暴雨成灾预测模型进行预测数据如表 6。

Table 6. Rainfall prediction
表 6. 降雨量预测

年份	预测降雨量
2025	0.999099
2026	0.987694
2027	1.010445
2028	1.014259
2029	1.010462
2030	0.999057
2031	1.021808
2032	1.025621
2033	1.021824
2034	1.010419

根据预测结果，未来十年(2025 至 2034 年)降水量呈现波动趋势，大多数年份的降水量与 2018 年接近，整体较为稳定。2026 年降水量可能下降至 0.987694，略有减少；而 2027 年和 2028 年则分别略有增长至 1.010445 和 1.014259。到 2031 年和 2032 年，降水量进一步增加至 1.021808 和 1.025621，这表明该阶段降水量可能受到气候变化或其他环境因素的影响，逐步增强。

5.3.3. 结果分析

本文使用逻辑回归模型和 Prophet 时间序列模型对暴雨成灾进行预测。逻辑回归模型通过 Logit > 0.79 判断暴雨成灾，其均方误差(MSE)为 0.05，平均绝对误差(MAE)为 0.07，分类准确率为 85%。Prophet 模型的 MSE 为 0.03，MAE 为 0.05，均方根误差(RMSE)为 0.17。不确定性评估显示，逻辑回归模型的 95%置信区间宽度为±0.05，预测概率的标准差为 0.03；Prophet 模型的 80%预测区间平均宽度为 0.20，95%预

测区间平均宽度为0.30，蒙特卡洛模拟结果表明95%的预测值落在0.95至1.05之间。模型的局限性包括数据量有限可能导致的过拟合，数据质量影响模型性能，以及模型假设可能不完全符合实际情况。未来研究将进行数据增强、模型融合、特征工程优化和超参数调优，以提高模型的泛化能力和预测准确性。采用随机森林进行拟合，自然灾害占极少样本的总量，模型倾向于将自然灾害的样本判别成无灾害情况，这是不平衡学习导致的。未来可以尝试采用过采样处理(SMOTE)，随机欠采样处理等以增加模型性能。

5.4. 自然/人文地理交互分析

5.4.1. AHP 模型

AHP层次分析法由学者萨蒂提出，是一种通过分解不同元素，得到不同目标、准则和方案的层次结构，并对各个因素赋予权重的方法[4]。本论文采用AHP层次分析法，建立一个用于评价人口、GDP、耕地率、绿地率等对中国土地利用变化影响的模型。该模型通过定性和定量分析，将自然因素(如耕地率、绿地率)、社会经济因素(如人口密度、人均GDP)与土地利用变化的效应相结合。选取人口、GDP数据集、土地利用类型，即数据集4、数据集5 [5] [6]和数据集6 [7]-[11]。

5.4.2. 模型建立

1) 评价体系的建立

我们选择人均GDP 、耕地率、绿地率、人均绿地覆盖率等，系统性地分析不同区域在人口、经济、土地利用和生态环境方面的关系，通过这些指标建立如图所示的指标体系，

2) 构建评价矩阵

为了描述不同区域的城市化和可持续发展的指标采用层次分析法数学模型进行评价。在层次分析法中，构建判断矩阵并计算各指标的权重是至关重要的步骤。它通过对决策目标下各个影响因素的重要性进行比较，通常基于理论分析、经验积累以及专家评分等方法来确定各因素的相对权重。

评价矩阵A是一个$n*n$的矩阵，其中a_{ij}表示第i个指标相对于第j个指标的重要性。评价矩阵是根据专家的对比评价来构建的。

矩阵形式：AHP (层次分析法)是一种常用于多标准决策的分析方法，通过构建判断矩阵，并利用特征向量来确定每个指标的权重，从而为每项决策生成综合评分。具体的实现流程如下：

$$A=\begin{bmatrix} 1 & a_{12} & a_{13} & \cdots & a_{1n} \\ \frac{1}{a_{12}} & 1 & a_{23} & \cdots & a_{2n} \\ \frac{1}{a_{13}} & \frac{1}{a_{23}} & 1 & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{a_{1n}} & \frac{1}{a_{2n}} & \frac{1}{a_{3n}} & \cdots & 1 \end{bmatrix}$$

其中，a_{ij}表示第i个指标相对于第j个指标的重要性，若$a_{ij}>1$，说明第i个指标比第j个更重要。根据土地类型、人均GDP，种植效率，人均绿地覆盖率本文的专家评价矩阵构建如表7所示。

Table 7. Expert evaluation matrix
表7. 专家评价矩阵

	耕地率	绿地率	人均GDP	种植效率	人均绿地覆盖率
耕地率	1	3	5	7	12

续表

绿地率	1/3	1	2	4	1/2
人均 GDP	1/5	1/2	1	3	1/3
种植效率	1/7	1/4	1/3	1	1/5
人均绿地覆盖率	1/12	2	3	5	1

3) 计算权重

为了得到每个指标的权重，需要计算评价矩阵的特征向量。步骤如下：计算评价矩阵 A 的特征向量 W：

$$AW = \lambda_{\max} W \tag{18}$$

其中，$\lambda_{\max}$ 是矩阵 A 的最大特征值，c 是对应的特征向量，经过归一化处理后，特征向量 W 就是每个指标的权重：

$$W_i = \frac{W_i}{\sum_{i=1}^{n} W_i} \tag{19}$$

计算得各权重如表 8 所示。

Table 8. Rainfall prediction
表 8. 降雨量预测

权重名称	权值
指标耕地率的权重	0.1604
指标绿地率的权重	0.1604
指标人均 GDP 的权重	0.0974
指标种植效率的权重	0.0463
指标人均绿地覆盖率的权重	0.2577

4) 计算一致性比率

为了检查专家评价矩阵的一致性，首先需要计算一致性指标 CI 和一致性比率 CR。一致性指标 CI：

$$CI = \frac{\lambda_{\max} - n}{n-1} \tag{20}$$

一致性比率 CR：

$$CR = \frac{CI}{RI} \tag{21}$$

其中，RI 是随机一致性指标，随矩阵阶数 n 不同，取对应的值。如果 $CR < 0.1$，则认为评价矩阵具有满意的一致性。一致性比率 CR:0.0178，评价矩阵的一致性良好，可以使用计算的权重。

5) 计算总得分

将权重 wi 和每个指标的评分 S；相乘，然后相加，得到每个对象的总得分 T：

$$T = \sum_{i=1}^{n} W_i S_i \tag{22}$$

其中，W_i 是第 i 个指标的权重，S_i 是第 i 个指标的评分。四个地区的 AHP 分数如图 7 所示。

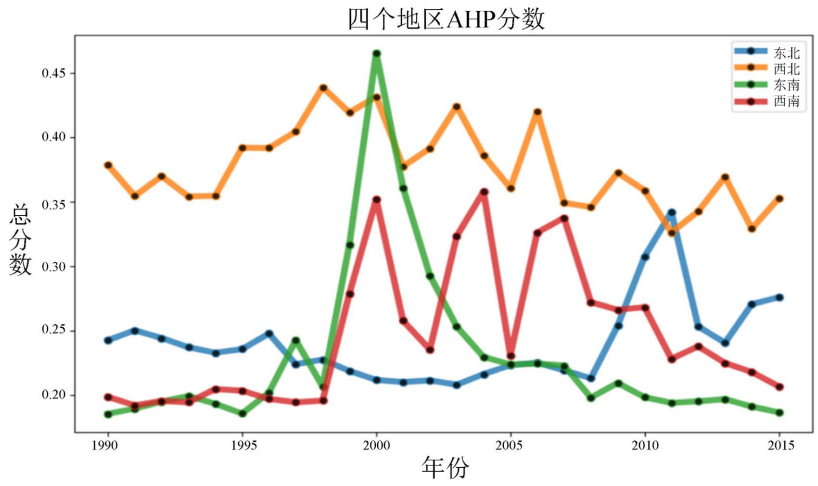

Figure 7. AHP score chart for four regions
图 7. 四个地区 AHP 分数图

秦岭淮河和胡焕庸线将中国划分为东南西北四个地区通过层次分析法得出四个地区的 AHP 分数验证可知，东北地区、西北地区在 2010 年后 AHP 分数较高，即城市化可持续发展评价较高。

6. 结论与展望

本文通过整合和分析地理大数据，深入挖掘了中国地理系统多要素时空特征与灾害预测之间的关联机制。研究发现，降水量、土地覆被类型和土地利用变化等要素对极端天气的形成具有重要影响。未来研究将聚焦于模型优化与改进，包括数据增强、模型融合、特征工程优化和超参数调优，以提高模型的泛化能力和预测准确性。针对不平衡学习问题，将采用过采样和随机欠采样技术，提升模型对自然灾害样本的识别能力。此外研究方向也可拓展至多源数据融合，整合卫星遥感、气象和社会经济数据，增强模型预测精度。同时，探索深度学习方法，如卷积神经网络(CNN)和长短期记忆网络(LSTM)，以处理高维地理数据和时间序列数据。跨学科的研究结合了地理学、气象学、生态学和社会学等多学科方法，深入探讨地理系统多要素的相互作用对人类社会进步提供了重要保障。

参考文献

[1] Han, J. and Miao, C. (2022) A New Daily Gridded Precipitation Dataset for the Chinese Mainland Based on Gauge Observations. *Earth System Science Data*, **15**, 3147-3161. https://doi.org/10.5194/essd-2022-373

[2] 余振, Philippe Ciais, 朴世龙, 等. 1900-2019 年中国土地利用和覆盖变化数据集[EB/OL]. https://doi.org/10.12199/nesdc.ecodb.pa.2022.11, 2024-10-02.

[3] Yu, Z., Ciais, P., Piao, S., Houghton, R.A., Lu, C., Tian, H., Agathokleous, E., Kattel, G.R., Sitch, S., Goll, D., Yue, X., Walker, A., Friedlingstein, P., Jain, A. K., Liu, S. and Zhou, G. (2022) Forest Expansion Dominates China's Land Carbon Sink since 1980. *Nature Communications*, **13**, Article No. 5374. https://doi.org/10.1038/s41467-022-32961-2

[4] 李泽京, 王勇, 王一, 等. 基于层次分析法的某煤矿水文地质类型评价[J]. 地下水, 2024, 46(5): 24-26.

[5] 王灿, 王嘉琛. 中国历史人口空间分布公里网格数据集(1990-2015 逐年) [EB/OL]. 国家青藏高原数据中心. https://doi.org/10.12078/2017121101, 2024-09-10.

[6] 徐新良. 中国人口空间分布公里网格数据集[EB/OL]. 资源环境科学数据注册与出版系统. http://www.resdc.cn/, 2024-10-08.

[7] 王灿, 王嘉琛. 中国历史 GDP 空间分布公里网格数据集(1990-2015) [EB/OL]. 国家青藏高原数据中心. https://doi.org/10.12078/2017121102, 2024-10-13.

[8] 徐新良. 中国 GDP 空间分布公里网格数据集[EB/OL]. 资源环境科学数据注册与出版系统. http://www.resdc.cn/, 2024-10-15.

[9] Liu, H., Jiang, D., Yang, X. and Luo, C. (2005) Spatialization Approach to 1 km Grid GDP Supported by Remote Sensing. *Geographic Information Science*, **7**, 120-123.

[10] 黄莹, 包安明, 陈曦, 刘海隆, 杨光华. 基于绿洲土地利用的区域GDP公里格网化研究[J]. 冰川冻土, 2009, 31(1): 158-165.

[11] Yi, L., Xiong, L. and Yang, X. (2006) Method of Pixelizing GDP Data Based on the GIS. *J. Gansu Sci*, **18**, 54-58.