****  **** 

**中国研究生创新实践系列大赛**

**“华为杯”第二十一届中国研究生**

**数学建模竞赛**

|  |  |
| --- | --- |
| **学 校** | **同济大学** |
| **参赛队号** | **24102470100** |
| **队员姓名** | 1. **郑浩** |
| 1. **邹锐卓** |
| 1. **江俊杰** |

**中国研究生创新实践系列大赛**

**“华为杯”第二十一届中国研究生**

**数学建模竞赛**

题 目： ­­­­\_\_\_\_\_\_\_\_\_\_\_大数据驱动的地理综合问题\_\_\_\_\_\_\_\_\_\_

摘 要：小组结合地理相关数据集的具体类型、格式，运用Python语言及相关工具，对大数据驱动的地理综合问题展开探究。首先对全国历史降水量、各类型土地面积等变量构建描述性统计方法，采用平均数、变化率等统计指标和折线图、热力图等图表形式进行分析；其次分别建立一元/多元线性回归模型，探究降水量与海拔、坡度之间的关系，以及极端天气事件与降水量、海拔、坡度、气温等参数之间的关系，进行相关性检验，通过损失函数等指标进行有效性评估，并结合随机森林模型对未来中国应对暴雨灾害较为脆弱的地区进行预测；最后结合之前所得结论，对数据进行简化和综合，描述中国土地利用变化的特征与结构。

关键词：大数据，地理综合，线性回归，随机森林

大数据驱动的地理综合问题

一、引言

地理系统是自然、人文多要素综合作用的复杂巨系统[1-2]，其复杂性体现在要素众多、交互作用复杂以及时空尺度多变等方面。为了更好地理解地理系统，地理学家们长期致力于地理综合的研究，利用不同的方法来揭示地理系统的整体特征和演化规律[3]。随着对地观测体系的快速发展，近年来积累了海量的对地观测数据。如何利用大数据手段对地理系统进行综合分析，探索全球气候变化下中国地理环境的演化，成为了当前地球科学研究的关键问题[5-6]。

本研究将聚焦于中国地理环境的时空演化，以降水量和土地利用/土地覆被类型为例，利用大数据分析方法，探讨这两个变量在1990-2020年间的时空演化特征，以及地形、气候和土地利用对极端天气形成的影响和暴雨成灾的脆弱区域预测。研究以中国地理系统的关键参数指标（如降水量、海拔、坡度、土地类型等）为研究对象，选取Python作为编程语言，利用相关库和工具进行数据分析和建模，对地理综合问题展开深入探讨。首先对全国历史降水量、各类土地面积等变量进行描述性统计分析，采用诸如平均数、变化率等统计指标，并通过折线图、热力图等图表展示分析结果。其次，分别建立一元和多元线性回归模型，以深入探究极端天气事件与降水量、海拔、坡度、气温等一系列参数之间的相互关系，并进行相关性检验、有效性评估指标以对模型进行验证。此外还将结合随机森林、XBoost等模型方法，对未来中国在面对暴雨灾害时较为脆弱的地区进行预测。最后，在以上分析的基础上进一步简化与综合数据，精简描述中国土地利用变化的特征与结构。

本研究的意义在于：（1）通过分析降水量和土地利用/土地覆被类型的时空演化特征，为理解中国地理环境的变化趋势提供科学依据；（2）建立地形-气候相互作用的数学模型，阐明地形在极端天气形成过程中的作用机制，为极端天气预报提供理论基础；（3）预测未来中国暴雨成灾的脆弱区域，为防灾减灾工作提供科学指导。研究成果将以论文的形式展现，并以图表等数据可视化的方式呈现研究结果。

二、建模思路

2.1 描述和总结降水量和土地利用/土地覆被的时空演化特征

首先需要选取相关数据集，为变量（降水量和土地利用/土地覆被）分别构建一套描述性统计方法，用1~3个较为简洁的统计指标或统计图表，对这两个变量在1990~2020年间中国范围内的时空演化特征进行描述和总结。

对于降水量数据，小组使用Python的netCDF4库读取1961-2022年中国大陆0.25°逐日降水数据集，并选取出1990~2020年的TIFF格式数据。原始数据的格式为一个三维数组，各维度变量分别为时间、经度、纬度，数组中的每个元素代表0.25°×0.25°区域（下称单位区域，其经纬度分别记为单位经度ula和单位纬度ulo）内的日降水量。小组选用3组统计指标对降水量的时空演化情况进行描述，在1990~2020的时间范围内，分别为全国各单位区域的历史平均降水量图（以热力图形式表示）、中国年均总降水量变化曲线（以折线图形式表示）以及中国历年月均降水量变化特征（以箱形图形式表示）。各指标处理和图形绘制的思路如下：

1. 将各单位区域的原始降水量数据对时间求平均值

其中, , 分别为经度（单位：ula）、纬度（单位：ulo）和时间（单位：天）,为1990至2020年的天数，即得各单位区域在目标时间范围内的平均值。

1. 求原始降水量数据的各单位区域之和，得到全国降水量的日平均值

其中, , 分别为经度（单位：ula）、纬度（单位：ulo）和时间（单位：天）,和分别为数据集地图中单位经度和单位纬度的数量。对日平均值乘以一年的天数，即可得到各单位区域在目标时间范围内的年平均值。

1. 对于全国降水量的日平均值，对每一年中对应的月份取平均，即可得到中国历年降水量月平均值。

对于土地利用/土地覆被数据，小组读取了1900-2019年中国0.5°土地利用和覆盖变化数据集，选取出1990~2020年的TIFF格式数据，

三、代码实现

四、运行结果

五、总结

参考文献

[编号] 卷不动的程序猿,机器学习之线性回归算法Linear Regression（Python代码实现）,https://blog.csdn.net/qq\_41750911/article/details/124883520,2024.9.23.