2024 年中国研究生数学建模竞赛 D 题

大数据驱动的地理综合问题

地理系统是自然、人文多要素综合作用的复杂巨系统[1-2]，地理学家常用地理综合的方式对地理系统进行主导特征的表达[3]。如以三大阶梯概括中国的地形特征，以秦岭—淮河一线和其它地理区划的方式揭示中国气温、降水、植被、土壤及生态环境在水平和垂直方向上的地带性与非地带性规律，利用胡焕庸线[4]、T 型开发结构等描绘我国人口、社会和经济发展的总体格局。这些方法早期以宏观结构和定性分析为主体，对我国生态保护、社会经济发展和国家安全保障起到了巨大的支撑作用。伴随着对地观测体系的快速发展，当前已经积累了巨量的对地观测数据。如何利用大数据的手段对地理系统进行综合[5-6]，探索全球气候变化下中国地理环境的演化，是当前地球科学研究的关键问题。

请利用“附件数据”一栏中列出的描述全球和中国地理不同方面特征的数据集，回答以下问题：

1. 在众多描述地理环境的变量中，一些简单的指标背后蕴藏了深厚的内涵，对人类的生存发展具有重大深远的影响，如大气中二氧化碳的浓度、全球年平均气温等。降水量是一个连续变化的变量，而土地利用/土地覆被类型则是一个存在突变和离散分布的变量。同时，它们都具有时空分布不均匀的特征。请从附件数据中选取相关数据集，为这两个变量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。
2. 近年来，以暴雨为代表的极端天气事件对人类的生产生活造成了越来越难以忽视的影响。请结合附件中所给的数据，建立数学模型，说明地形-气候相互作用在极端天气形成过程中的作用。
3. 降雨、地形和土地利用对于暴雨等极端天气灾害的形成都具有不可忽视的影响。这其中，降雨的时空变异性和不可控性都最强；土地利用作为自然条件和人类活动的综合结果，虽然也随时空演化，但具有一定可控性；地形是最为稳定、不易改变的因素。请考虑第 2 问所反映的从“暴雨”到“灾害”中上述三方面因素的角色及其交互作用，确定暴雨成灾的临界条件；并结合第 1 问中降雨量和土地利用/土地覆被变化的历史时空演化特征，对 2025~2035 年间中国境内应对暴雨灾害能力最为脆弱的地区进行预测。请以地图的形式呈现你们的预测结果。
4. 在中国级别的尺度上，描述自然地理特征的地形可以概括为“三级阶梯”，而降水中具有标志性意义的“800mm 等降水量线”则与区分我国南北方的“秦岭—淮河”一线大体重合；描述人文地理特征的人口分布及其社会经济活动

总量等指标，则被由连接黑龙江黑河与云南腾冲的“胡焕庸线”清晰地划分成东密西疏的两部分。那么，对于自然地理和人文地理交汇点的土地利用/土地覆被情况，结合其在前三问中描述、估计和预测任务中的“特性”，利用地理大数据，建立相应的数学模型，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。从准确性和有用性两个方面解释验证你们的总结。

名词解释

1. 地理综合：指采用系统性的思维，综合运用各种地理信息和分析方法，对地理系统的整体结构、功能和演变规律进行整体性认知和描述的过程。地理综合通常体现在综合自然要素、综合人文要素、综合时空尺度、综合定性定量分析和综合基础理论与应用几个方面。
2. 复杂巨系统：由大量相互联系、相互作用的要素组成的超大型系统。复杂巨系统通常由成千上万个要素组成，涉及的子系统及层次众多。系统内部各要素之间存在着错综复杂的相互联系和反馈机制，随时间演化且难以用简单的因果关系描述。此外，复杂巨系统还具有非线性、开放性、自组织性等特征。
3. 三大阶梯：我国的地形格局西高东低，可以形象地概括为“三大阶梯”，其中第一级阶梯为青藏高原，平均海拔在 4000 米以上，被称为“世界屋脊”；青藏高原以东是第二级阶梯，以大兴安岭—太行山—雪峰山为界，海拔大多在 1000-2000 米之间；第二级阶梯以东是第三级阶梯，大部分地区海拔在 500 米以下。
4. 土地利用/土地覆被：土地利用（land use）与土地覆被（land cover）是一对既有联系又有区别的概念。土地利用描述了“人类如何利用土地资源”，其结果可以是农业用地、工业用地、交通用地、居住用地等。土地覆被描述了“陆地表面是何种状态”，如各类作物、森林、草地、房屋、水泥及沥青路面等。一方面，土地利用是土地覆被发生改变的主要原因；另一方面，土地覆被也为土地利用提供了前提条件或制约。
5. 描述性统计方法：在对数据不持有预设立场的情况下，使用数字、图表等方式，对原始数据进行定量总结、概括的方法。描述性统计方法一般关注数据的集中趋势、离散程度、分布形态等特征，经典的描述性统计指标有平均数、方差、折线图、散点图、频率分布直方图等。与之相对应的是推断性统计方法，即根据预设的立场（称为假设），对数据进行特定的统计操作（称为假设检验），从而证实或推翻预设立场的方法。
6. 等降水量线：将地球表面年降水量（包括降雨、降雪、降冰雹等）相等的点连接而成的线。在我国，800mm 等降水量线不仅与“秦岭—淮河”这一南北地理分界线重合，也与一月份 0℃等温线、水田与旱地的分界线、水稻和小麦种植分界线、亚热带与暖温带的分界线、湿润与半湿润的分界线大体重合，具有重要的地理意义。

人工智能产品辅助答题规范

以大模型为代表的人工智能产品（诸如 ChatGPT、Perplexity、文心一言、通义千问等）已经在人们的生产生活不同方面崭露头角。如果你们在回答上述问题时采用了人工智能产品，请遵循以下原则：

1. 把人工智能产品作为答题的辅助工具，而非主导手段。人工智能产品可以在信息搜集、开拓思路和工具学习等方面助一臂之力，但是不可以替代你们的独立思考。

2. 对人工智能产品的输出应当先理解，再利用。最终呈现在解答中的内容，应当是你们自己的语言。如果在查重检测中发现你们论文使用人工智能产品与其他参赛队的输出内容雷同，将被判为违规，并按竞赛规定受到处罚。
3. 在使用人工智能产品辅助论文写作中，正文和数学模型及公式引用出处应是正式发表的文献或输入论文所提供的网址可在网上查询到的内容，而非引用人工智能产品得到的内容或结果。
4. 披露对人工智能产品的使用情况。答题过程中，如果使用了人工智能技术，请在正文最后以附录形式披露所采用的人工智能产品、提供的输入及对输出的后续处理策略，内容包括但不限于算法组合采用的开发框架、开源软件，算法逻辑中的技术路线、假设条件、参数与超参数等。

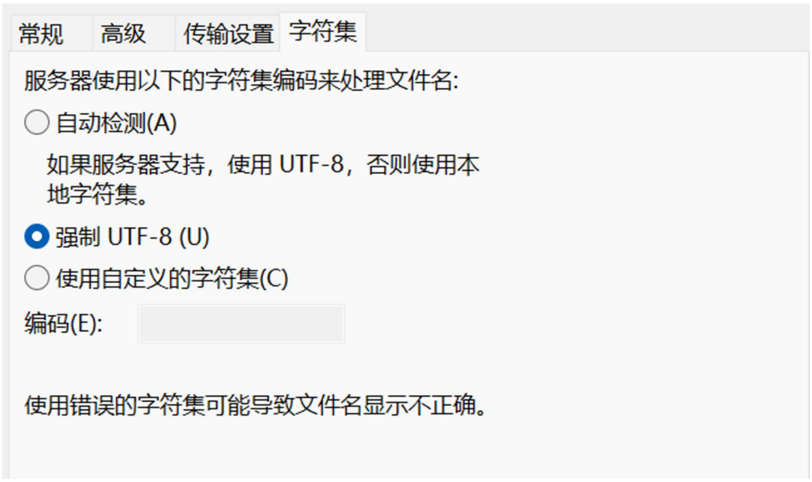
附件数据

由于本题的附带数据较大（6 个数据集，共 6.24GB），参赛选手可从赛前 24 小时（北京时间 2024 年 9 月 20 日上午 8:00）起，至竞赛结束（北京时间 2024 年 9 月 25 日中午 12:00）止，下载加密过的数据集压缩包。数据下载时间充裕，各位选手可错峰下载，也可在下载完后共享给其他选手使用。

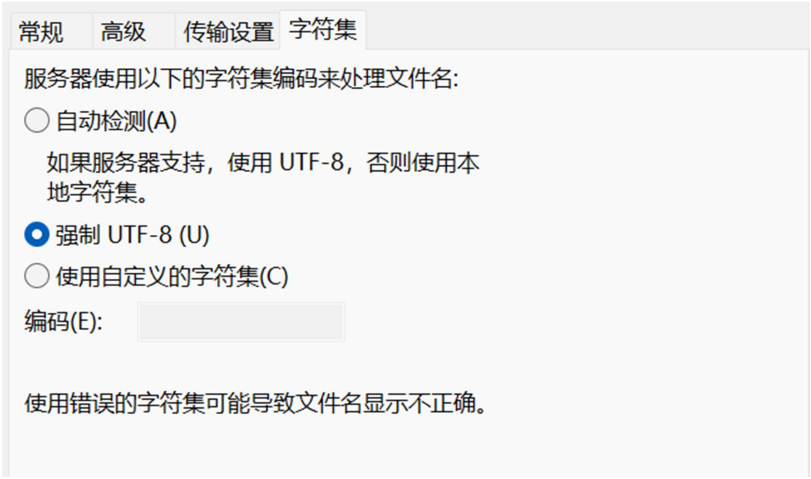
数据集的解压缩密码见后文各数据集介绍中的黄色高亮部分。

为保障选手能够顺利下载赛题附件数据，大赛设置 4 个 FTP 下载站点和 1 个百度网盘下载地址。FTP 下载方式推荐使用 FileZilla 软件。FileZilla 是一款开源免费 FTP 客户端软件（官方网站：https://filezilla-project.org/），其相关设置如下：依次点击文件→站点管理器→新站点，常规→主机，然后填写相关配置信息。以下四套配置信息，任选一套填写即可。

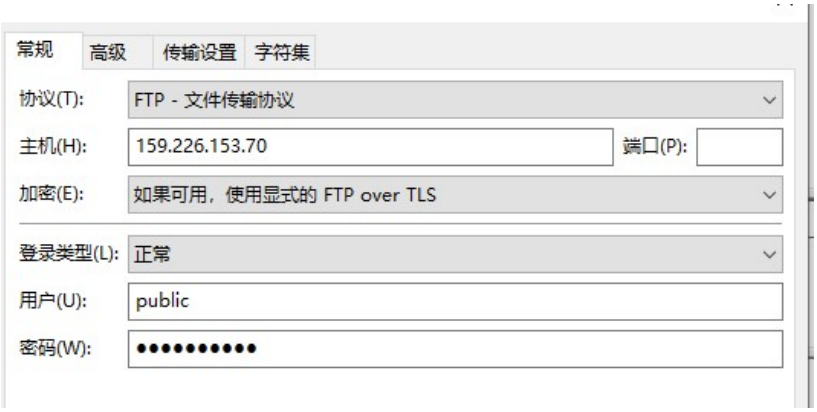
配置信息 1：
主机(H)：159.226.115.140
加密(E)：“如果可用，使用显式的 FTP over TLS”
登录类型(L)：匿名
在“高级”标签页中，默认远程目录(E)：/PullDir/sihsjkd_240822_140827_0575
字符集选择“强制 UTF-8(U)”

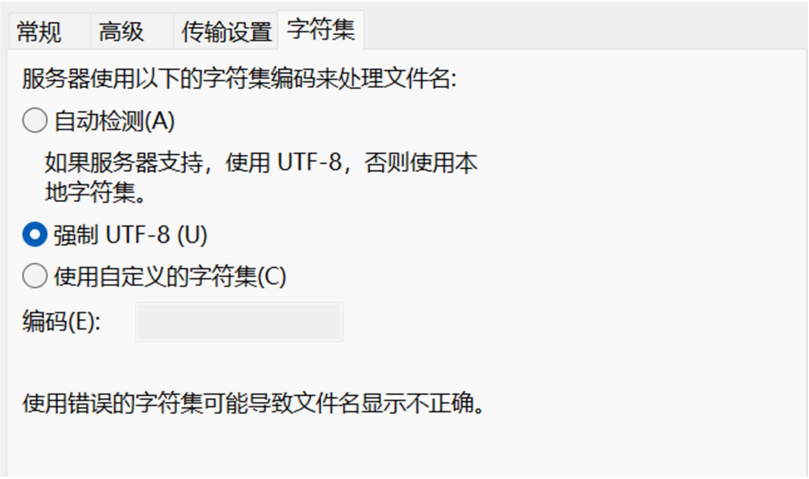


配置信息 2：
主机(H)：222.192.7.77
加密(E)：“要求隐式的 FTP over TLS”
登录类型(L)：匿名
字符集选择“强制 UTF-8(U)”



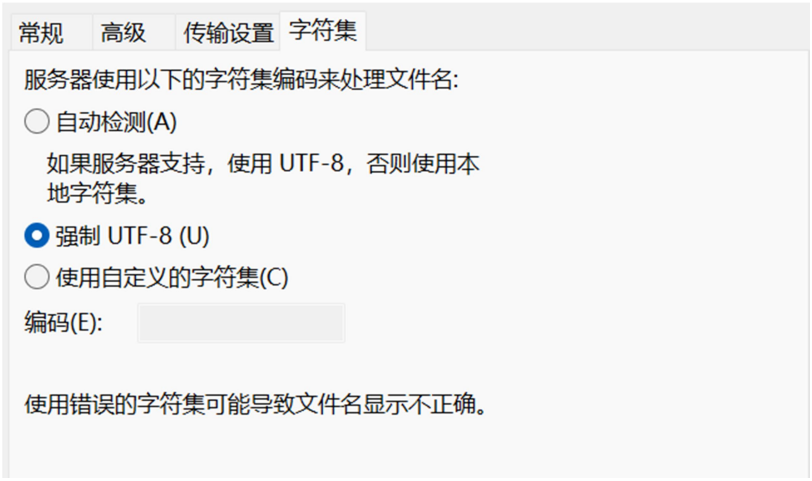
配置信息 3：
主机(H)：159.226.153.70
加密(E)：“如果可用，使用显式的 FTP over TLS”
用户(U)：public
密码(W)：iswc712100
字符集选择“强制 UTF-8(U)”





配置信息 4：
主机(H)：210.77.90.99
加密(E)：“如果可用，使用显式的 FTP over TLS”

用户(U)：data2024
密码(W)：data2024@SCSODC
字符集选择“强制 UTF-8(U)”



百度网盘下载
链接：https://pan.baidu.com/s/1C1IlX_5NRHcLxuDx485Vwg?pwd=twk7
提取码：twk7

如遇技术问题，可联系 kutukutuku8989@163.com。

下文涉及对角度的描述中，1°、1′、1″分别表示角度 1 度、1 分、1 秒，1°= 60′ = 3600″ 。

数据集涉及 GeoTIFF 和 NetCDF 两种格式的文件。这两种文件都可以采用 Python 和 R 等开源编程软件处理，也可以采用地理数据处理软件如 ArcGIS、GeoScene、SuperMap、QGIS 等进行处理。其中 QGIS 是免费开源的地理信息处理软件，任何人都可下载使用。

QGIS 官方网站：https://qgis.org/
QGIS 教程：https://docs.qgis.org/3.34/en/docs/training_manual/index.html
https://www.osgeo.cn/qgis-tutorial/index.html（中文教程）

1. 中国数字高程图（1km）

（数据集 1，138MB，解压缩密码：chvmg8）

该数据集包含两种采用两种不同坐标系的数据。解压缩后，文件夹 Albers_105

内为采用正轴割圆锥等面积投影（Albers Conical Equal Area Projection）的数据，文件夹 Geo 内为采用 WGS84 地理坐标系的数据。在 Geo 文件夹内 TIFF 子文件夹下有文件 chinadem_geo.tif，是一幅 GeoTIFF 图像。图像中每个像素的大小为 1km×1km，像素的值代表了地表对应位置的海拔高度，单位为米。更多详情请参见随数据文件一同下载的说明文件。

如使用本数据集，须在赛题的解答中按以下方式进行引用：

[1] 汤国安. (2019). 中国数字高程图（1KM）. 国家青藏高原数据中心. [Tang, G. (2019). Digital elevation model of China (1KM). National Tibetan Plateau / Third Pole Environment Data Center.]

2. 中国 0.1°近地表气温数据集（1979-2018 年）

（数据集 2，4.27GB，解压缩密码：tJpFHN）

该数据集包含全国 1979-2018 年每日的气温平均值的分布。每一年的数据为一个文件夹，内含当年 1 月 1 日至 12 月 31 日每一天的气温分布，以 GeoTIFF 图像保存。图像中每个像素的大小为 0.1°×0.1°，像素的值代表了地表对应位置的气温，单位为摄氏度。更多详情请参见随数据文件一同下载的说明文件。

如使用本数据集，须在赛题的解答中按以下方式进行引用：

[1] Fang, S., Mao, K., Xia, X., Wang, P., Shi, J., M. Bateni, S., Xu, T., Cao, M., & Heggy, E. (2021). A Daily near-surface Air Temperature Dataset for China from 1979 - 2018 (Version 1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5502275

3. 中国大陆 0.25°逐日降水数据集（1961-2022 年）

（数据集 3，288MB，解压缩密码：p7wDju）

该数据集是一个 NetCDF 文件。文件中包含了 22645 层图像，每一层图像对应从 1961 年 1 月 1 日至 2022 年 12 月 31 日每一天的降水量分布（中国台湾省及南海地区数据暂缺）。每一层图像中，像素大小为 0.25°×0.25°，像素的值代表了地表对应位置的降水量，单位为毫米。更多详情请参见随数据文件一同下载的说明文件。

如使用本数据集，须在赛题的解答中按以下方式进行引用：

[1] Han, J., Miao, C. (2022). A new daily gridded precipitation dataset for the Chinese mainland based on gauge observations. figshare. Dataset. https://doi.org/10.6084/ m9.figshare.21432123.v4

4. 中国 0.5°土地利用和覆盖变化数据集（1900-2019 年）

（数据集 4，10.0MB，解压缩密码：GkWwGa）

该数据集包含 1900-2019 年耕地、林地、草地、灌木丛、湿地五种土地覆被类型在中国的分布情况，以 GeoTIFF 图像格式保存。每一类覆被类型的文件分别以覆被类型的英文名词开头，以年份结尾。图像中每个像素的大小为 0.5°×0.5°，像素的值在 0~1 之间，表示像素中该覆被类型所占面积比例。更多详情请参见随数据文件一同下载的说明文件。

如使用本数据集，须在赛题的解答中按以下方式进行引用：

[1] 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[DS/OL]. 国家生态科学数据中心, 2022.
https://doi.org/10.12199/nesdc. ecodb.pa.2022.11
https://cstr.cn/15732.11.nesdc.ecodb.pa.2022.11

[2] 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., & Zhou, G. (2022). Forest expansion dominates China's land carbon sink since 1980. *Nature Communications, 13*(1), 5374. https://doi.org/10.1038/s41467-022-32961-2

5. 中国大陆 1km 逐年历史人口空间分布公里网格数据集（1990-2015 年）
（数据集 5，818MB，解压缩密码：3NSzbY）
该数据集包含 26 个文件夹，每个文件夹内包含了某一年中国人口的空间分布，数据以一幅 GeoTIFF 图像格式保存。图像中每个像素的大小为 1 千米×1 千米，像素的值为地表对应范围内的人口数估计值（中国台湾省及南海地区数据暂缺）。更多详情请参见随数据文件一同下载的说明文件。

如使用本数据集，须在赛题的解答中按以下方式进行引用：

[1] 王灿, 王嘉琛. (2022). 中国历史人口空间分布公里网格数据集（1990-2015 逐年）. 国家青藏高原数据中心. https://doi.org/10.12078/2017121101 [Wang, C., Wang, J. (2022). Kilometer grid dataset of China's historical population spatial distribution (1990-2015). National Tibetan Plateau / Third Pole Environment Data Center. https://doi.org/10.12078/2017121101]

[2] 徐新良. (2017). 中国人口空间分布公里网格数据集.资源环境科学数据注册与出版系统(http://www.resdc.cn/DOI). DOI:10.12078/2017121101

6. 中国大陆 1km 逐年历史 GDP 空间分布公里网格数据集（1990-2015 年）
（数据集 6，759MB，解压缩密码：aEKevB）
该数据集包含 1990-2015 年每一年中国 GDP 的空间分布，每一年的数据以一幅 GeoTIFF 图像格式保存。图像中每个像素的大小为 1 千米×1 千米，像素的值为地表对应范围内的 GDP，单位为万元人民币（中国台湾省及南海地区数据暂缺）。更多详情请参见随数据文件一同下载的说明文件。

如使用本数据集，须在赛题的解答中按以下方式进行引用：

[1] 王灿, 王嘉琛. (2022). 中国历史 GDP 空间分布公里网格数据集(1990-2015). 国家青藏高原数据中心. https://doi.org/10.12078/2017121102 [Wang, C., Wang, J. (2022). Kilometer grid dataset of China's historical GDP spatial distribution (1990-2015). National Tibetan Plateau / Third Pole Environment Data Center. https://doi.org/10.12078/2017121102]

[2] 徐新良. (2017). 中国 GDP 空间分布公里网格数据集. 资源环境科学数据注册与出版系统(http://www.resdc.cn/DOI). DOI:10.12078/2017121102

[3] Liu, H., Jiang, D., Yang, X., & Luo, C. (2005). Spatialization approach to 1 km grid GDP supported by remote sensing. Geo-Inf. Sci, 7, 120-123.

[4] 黄莹, 包安明, 陈曦, 刘海隆, & 杨光华. (2009). 基于绿洲土地利用的区域 GDP 公里格网化研究. 冰川冻土, (1), 158-165.
[5] Yi, L., Xiong, L., & Yang, X. (2006). Method of pixelizing GDP data based on the GIS. J. Gansu Sci, 18, 54-58.

参考文献

[1] 陈述彭. 地理系统与地理信息系统 [J]. 地理学报, 1991, 46(1): 1-7. https://doi.org/10.11821/xb199101001
[2] 彭书时, 朴世龙, 于家烁, 刘永稳, 汪涛, 朱高峰, 董金玮, 缪驰远. 地理系统模型研究进展[J]. 地理科学进展, 2018, 37(1): 109-120. https://doi.org/10.18306/dlkxjz.2018.01.012
[3] Peuquet, D. J. (1988). Representations of Geographic Space: Toward a Conceptual Synthesis. Annals of the Association of American Geographers, 78(3), 375-394. https://doi.org/https://doi.org/10.1111/j.1467-8306.1988.tb00214.x
[4] 胡焕庸. 中国人口之分布——附统计表与密度图 [J]. 地理学报, 1935, 2(2): 33-74 https://doi.org/10.11821/xb193502002
[5] Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. Dialogues in Human Geography, 3(3), 255-261. https://doi.org/10.1177/2043820613513121
[6] 闾国年, 周成虎, 林珲, 陈旻, 乐松山, 温永宁. 地理综合研究方法的发展与思考 [J]. 科学通报, 2021, 66(20): 2542-2554. https://doi.org/10.1360/TB-2020-0799