

| 前 言 |

地理学是一门研究地理要素或者地理综合体的空间分布规律、时间演变过程和区域特征的一门学科，也是一门研究人地关系的科学，尤其强调了人地相互关系的耦合，同时地理学还是一门经世致用的综合性学科。地理学的内涵和特征决定了该学科发展的终极目标是服务于国家宏观战略的制订，以促进人类生活环境的宜居和可持续发展及公共健康的全面提升。地理大数据的来源于不只限于空间对地观测数据，还包括地面、海洋、大气及人类行为等相关领域的的数据。地理大数据已经成为我们认识地球的新窗口，也将成为服务人类社会的新工具，充分挖掘地理大数据的价值、揭示其中蕴含的知识和规律，不仅将更好地服务于宏观决策，还可为推动学科的深入发展和重大科学新发现做出贡献，并将在促进全球可持续发展方面发挥重大作用。

世界卫生组织于 2020 年 3 月 11 日正式宣布新型冠状病毒肺炎（COVID-19）为“全球性流行病”。与新型冠状病毒肺炎在全球快速蔓延相对应的是我国新型冠状病毒肺炎疫情（以下简称新冠肺炎疫情）迅速得到了有效地控制，中国抗击疫情的显著成效为全球抗击疫情带来了希望，中国经验也值得全球借鉴。我国的成功除了归功于奋战在抗疫一线的广大医务工作者、疾控工作人员和各级基层工作人员外，许多地理学研究者的贡献也功不可没。如中国科学院周成虎院士团队收集整理全国疫情数据，即时制作和发布全国疫情分布图和新冠肺炎疫情的风险地图；该团队还利用新冠肺炎疫情大数据进行科学地分析，提出了区域防控政策建议；北京师范大学田怀玉联合国内外专家定量评估了我国新冠肺炎疫情的传播和控制效果，发现中国积极的交通管控措施对疫情的有效控制起到了关键作用，该成果已在 *Science* 上发表；北京师范大学傅伯杰院士团队在 *Science Bulletin* 向全球推介了中国抗疫的成功模式；兰州大学黄建平团队研发了世界首个新冠肺炎疫情全球预测系统。基于我们地学科研工作者的格局、视野和专业背景，再次验证了地理学大数据分析方法和技术可以为公共健康事业作出卓越的贡献。

今年是我主讲安全科学与工程专业《统计学》课程的第三个年头，参考了许多国内外的教材，其中计算机语言类书籍和专业需求不甚匹配，理论类专业书籍又略显晦涩，因此一直未能选到合适的教材，才有了撰写本书的念头。本书采用理论和实践相结合，深入浅出地系统讲解地理大数据在公共健康领域的基本知识

和基本技术，在内容设计上既有课堂教学的讲解部分，包括详细的理论与典型案例，又安排了大量学生动手实践的环节，双管齐下，可望大大激发学生们的课堂上的学习积极性与主动创造性，让学生在课堂上紧随老师的思维和节奏，从而高效掌握更多有用的知识和技能。同时结合一门计算机语言软件的专业训练，将有利于培养和提高学生们的科研素质。

本书共 8 章，主要包括基本概念和理论方法介绍、地理大数据的采集（GitHub、爬虫、公共健康数据采集）、大数据的预处理和编码、不同类型大数据的可视化基础与应用、时间序列分析（时序数据的分解技术、预测模型构建等）、大数据的空间分析、COVID-19 的综合实训（数据获取、疫情图制作、疫情动力模型的构建、管控措施对疫情防控的影响、气象要素对疫情传播的影响、疫情和空气质量间的关系解析）、极端温度与公共健康案例分析（高温热浪和城市公共健康、长江三角洲极端温度与人口死亡关系的探讨）。本书建议学时与编写人员如表 0-1。

表 0-1 建议学时与编写人员

章节	建议学时	编写人员
大数据和公共健康概述	2	李子悦、张朝
大数据的采集	2	韩继冲、李子悦、张朝
大数据预处理	2	梅晴航、张朝
R 语言的可视化工具	4	吴华清 、张朝
大数据的时间序列分析	4	李子悦、张朝
空间数据处理与分析	4	韩继冲、张朝
基于 COVID-19 的综合实训	18	李子悦、韩继冲、梅晴航、张朝
极端温度与公共健康案例研究	12	吴华清 、梅晴航、金月雄、张朝

本书是北京师范大学地理科学学部张朝团队共同编写的成果，在编写过程中我们得到了北京师范大学地理科学学部领导傅伯杰院士、宋长青部长、武建军教授，以及安全科学与工程学科汪明、张文生、曹红斌、孟耀斌等多位老师们的大力支持。在编写过程中我们也参阅了大量的相关资料，科学出版社在书稿的编写和修订过程中给予了宝贵意见。在此，我们一并表示衷心的感谢！

本书如有疏漏、不妥之处，恳请广大读者赐教指正。

张 朝

2021 年 2 月于北京

| 目 录 |

前言

第 1 章 大数据和公共健康概述	1
1.1 公共健康概述	1
1.2 公共健康风险管理	8
1.3 大数据概述	13
参考文献	27
第 2 章 大数据的采集	29
2.1 GitHub 与大数据	29
2.2 爬虫与大数据	35
2.3 公共健康领域数据的采集	62
第 3 章 大数据预处理	69
3.1 创建数据集	69
3.2 R 语言的数据清洗	88
第 4 章 R 语言的可视化工具	102
4.1 R 语言三大绘图系统概述	102
4.2 基础图表绘制	109
4.3 专业图表绘制	137
参考文献	162
第 5 章 大数据的时间序列分析	163
5.1 时间序列数据	163
5.2 时序的平滑和季节性分解	170
5.3 指数预测模型	176
5.4 ARIMA 预测模型	183
第 6 章 空间数据处理与分析	196
6.1 空间数据处理	196
6.2 空间数据分析	204
第 7 章 基于 COVID-19 的综合实训	217
7.1 COVID-19 数据的获取	218

7.2 对 COVID-19 进行基础绘图	226
7.3 探究 1: 动力学模型下的 COVID-19 传播	240
7.4 探究 2: 气象要素对 COVID-19 传播的影响	282
7.5 探究 3: COVID-19 各省疫情和空气质量关系案例	310
参考文献	320
第 8 章 极端温度与公共健康案例研究	321
8.1 极端温度和公共健康	321
8.2 长江三角洲地区极端温度与人口死亡关系案例研究	359
参考文献	379