参赛队号：（参赛队无须填写，参赛队号由大赛官网自动生成）

2024年（第十届）全国大学生统计建模大赛 参 赛 作 品

|  |  |
| --- | --- |
| 参赛学校： | 北京邮电大学 |
| 论文题目： | “碳”索未来，“智”控千里——大数据与人工智能视角下的中国碳排放统计研究与减排策略 |
| 参赛队员： | 杜泽宣、孙佳辰、谢博任 |
| 指导老师： | 李亚杰 |

“碳”索未来，“智”控千里——大数据与人工智能视角下的中国碳排放统计研究与减排策略

摘要

碳排放问题是制约可持续发展的关键瓶颈。因此，碳减排是碳中和目标达成工作中的重要组成部分。针对传统统计方法难以应对碳排放数据规模庞大、来源多样的问题，本文提出了结合大数据与神经网络的解决方案。

本文基于中国分部门核算碳排放清单数据和中国表观碳排放清单数据，采用SAS软件分析和人工筛查的方法选择了合适的变量。通过绘制不同年份的碳排放指标占比图，展现了以火电为主发电方法和以煤炭为主的供热方式在碳排放中的主要地位，描绘了生产和供应电力、蒸气和热水的碳排放占比随年份增长的趋势。

本文通过空间聚类分析模型，应用K-Means算法对空间加权后的碳排放特征进行聚类，将所有省份按碳排放特征和发展方向分为五组，并对碳排放数据的地域差异问题进行了总结。随后，本文使用人工智能方法对碳排放数据进行拟合，结果表明梯度提升回归模型在拟合方面具有较高的准确性和解释能力，而LSTM-CNN混合模型在拟合碳排放数据方面综合性能最佳。最后，本文使用ARIMA模型对碳排放数据进行预测，得出了我国碳排放量讲会逐年递增的结论。

本文的研究结论是：结合大数据与人工智能技术的碳排放统计与管理策略，能够为实现中国碳中和目标提供强有力的技术支撑和决策依据。

关键词 碳排放 人工智能 因子分析 空间聚类分析 CNN-LSTM

目录

[摘要 2](#_Toc166684508)

[表格与插图清单 1](#_Toc166684509)

[一、绪论 2](#_Toc166684510)

[（一）研究背景 2](#_Toc166684511)

[（二）政策解读 2](#_Toc166684512)

[（三）研究意义 4](#_Toc166684513)

[（四）研究方法和创新 4](#_Toc166684514)

[（五）文献综述 5](#_Toc166684515)

[二、数据处理 6](#_Toc166684516)

[（一）数据来源 6](#_Toc166684517)

[（二）数据预处理和数据描述 6](#_Toc166684518)

[三、模型构建和分析 9](#_Toc166684519)

[（一）模型假设 9](#_Toc166684520)

[（二）对碳排放数据的地域考察——空间聚类分析 10](#_Toc166684521)

[1.模型的引入 10](#_Toc166684522)

[2.空间权重矩阵 10](#_Toc166684523)

[3.K-Means聚类 10](#_Toc166684524)

[（三）对碳排放数据的拟合——机器学习和深度学习模型 12](#_Toc166684525)

[1. 机器学习模型 12](#_Toc166684526)

[2. 深度学习模型 12](#_Toc166684527)

[（四）对碳排放数据的预测——ARIMA模型 14](#_Toc166684528)

[四、结论和建议 15](#_Toc166684529)

[（一）研究结论 15](#_Toc166684530)

[（二）减排策略 16](#_Toc166684531)

[参考文献 17](#_Toc166684532)

[附录 19](#_Toc166684533)

[致谢 20](#_Toc166684534)

# 表格与插图清单

表1 空间聚类分析模型的聚类结果

表2 机器学习模型性能评估

表3 深度学习模型性能评估

表4 碳排放数据差分结果

表5 ARIMA参数表

图 1 关于我国碳排放相关政策的词云图

图 2中国碳排放指标图例

图 3 1997年中国碳排放指标占比

图 4 2010年中国碳排放指标占比

图 5 2021年中国碳排放指标占比

图 6 1997年和2021年全国各省二氧化碳排放量

图 7 差分后数据的ACF、PACF图

图 8 ARIMA模型对我国碳排放量的预测

“碳”索未来，“智”控千里——大数据与人工智能视角下的中国碳排放统计研究与减排策略

# 一、绪论

## （一）研究背景

随着全球经济的持续扩张与工业化进程的深化，中国作为全球经济体中的重要一员，其快速发展的同时也面临着巨大的环境压力，尤其是碳排放问题凸显，成为制约可持续发展的关键瓶颈。在应对全球气候变化的大潮中，中国不仅承担着自身减排的重任，亦是国际减排合作的重要参与者。一方面，中国碳排放数据的规模庞大、来源多样，传统统计方法难以实现全面、动态且精准的管理。大数据技术的应用，能够整合与分析跨领域、跨行业的排放数据，为碳排放的精准量化、动态监测提供强有力支持。另一方面，人工智能技术，尤其是机器学习与深度学习算法，能够挖掘数据背后的复杂关联，预测排放趋势，为制定高效、前瞻性的减排政策提供科学依据。社会经济的多元化发展与公众环保意识的觉醒，进一步催生了对智能化、精细化碳管理解决方案的需求。从工业生产的能效提升到交通系统的智慧化改造，从城市规划的绿色导向到能源结构的转型升级，大数据与人工智能的深度融合为实现碳排放的有效管控与经济社会的绿色发展开辟了新路径。

## （二）政策解读

随着全球气候变化挑战日益严峻，特别是在大数据与人工智能技术蓬勃发展的背景下，中国的一系列碳排放相关政策展现出鲜明的时代特征和创新路径。

**碳达峰碳中和目标引领下的政策框架。**2020年9月，中国宣布了2030年前碳达峰、2060年前碳中和的目标，为国家整体的低碳发展设定了清晰的时间表和路线图。这一目标成为了后续各项政策制定的基石，推动经济结构、能源结构的根本性调整。通过发布《中共中央 国务院关于完整准确全面贯彻新发展理念做好碳达峰碳中和工作的意见》和《2030年前碳达峰行动方案》等文件，构建了从宏观到微观、跨行业跨领域的政策体系，明确了碳减排的总体方向和具体任务。

**大数据与人工智能技术的融合应用。**实现“碳达峰、碳中和”意味着深刻的系统性变革，而人工智能作为科技革命的关键要素，在为低碳发展开辟新路径的过程中发挥着重要作用。政府鼓励运用大数据技术建立和完善碳排放监测与核算体系，实现碳排放数据的实时采集、精准分析与高效管理。这不仅提高了数据的透明度和准确性，也为制定更加科学合理的减排策略提供了坚实基础。人工智能技术被广泛应用于能源生产和消费的各个环节，如智能电网的优化调度、风能和太阳能发电预测、以及建筑能效管理等，有效提升了能源利用效率，促进了清洁能源的规模化应用。

**碳市场机制的创新与实践。**2021年，全国碳排放权交易市场正式启动运营，标志着中国碳定价机制的重大进展。该市场依托于大数据平台，实现了对重点排放单位的精准监管和高效交易，为控排企业提供了经济激励，推动了低成本减排路径的探索，全面改善了碳排放数据质量。

**产业政策与技术创新并重。**通过实施绿色制造工程，推动传统产业智能化、绿色化改造，大数据和AI技术在工艺优化、能耗管理等方面的应用，助力工业部门大幅降低碳排放强度。加快新能源汽车推广，利用智能交通系统优化交通流，减少拥堵和排放。同时，大数据分析支持充电基础设施的合理布局，提升电动车使用便利性。

**法规制度与国际合作的强化。**修订后的《环境保护法》、《碳排放权交易管理办法》等法律法规，为碳排放管理和减排行动提供了坚实的法律基础，确保政策执行的有效性和连续性。积极参与国际气候治理，分享中国在大数据与人工智能辅助碳管理方面的经验和技术，加强与全球伙伴在绿色技术、资金和市场机制等方面的交流合作。



图 1 关于我国碳排放相关政策的词云图

## （三）研究意义

理论意义。本文的理论意义主要体现在三个方面：首先，推动环境经济学与数据科学的融合：在大数据与人工智能技术的支持下，中国碳排放统计研究能够更精准地量化不同行业、区域乃至微观个体的碳足迹，促进环境经济学理论与现代信息技术的深度融合。这不仅丰富了碳排放计量方法论，也为其他环境问题的研究提供了跨学科的方法借鉴。其次，深化对碳排放机理的理解：通过大数据分析，可以揭示碳排放与经济发展、人口分布、能源结构等多因素之间的复杂关系，为理解碳排放的社会经济驱动机制提供实证基础。人工智能技术如机器学习和深度学习能进一步挖掘隐藏的规律和模式，提高预测碳排放趋势的准确性，为制定科学合理的减排路径提供理论支撑。最后，促进环境监测技术创新：结合物联网、卫星遥感等技术，大数据与AI能够实现碳排放的实时监测和智能预警，推动环境监测从被动响应向主动预防转变，为环境科学领域带来技术革新。

实践意义。本文的实践意义主要体现在三个方面：首先，提升碳排放管理效率：利用大数据技术整合多源数据，构建全国或区域级碳排放数据库，可为政府、企业和第三方机构提供准确、及时的碳排放信息，优化资源配置，提高碳排放管理和减排措施的实施效率。其次，促进低碳技术与产业创新：基于AI的碳排放预测模型能帮助企业识别减排潜力大的环节，引导资金和技术向低碳、零碳领域倾斜，加速清洁能源、节能技术和循环经济的发展，推动产业结构转型升级。最后，增强公众参与度与意识：通过数据可视化和交互式平台，大数据与AI能让碳排放数据更加透明化，提高公众对气候变化的认识和参与碳减排行动的积极性，形成全社会共同参与的良好氛围。

政策意义。本文的政策意义主要体现在两个方面：首先，精确的碳排放统计数据和智能分析结果能够为政府制定碳达峰、碳中和相关政策提供量化的决策支持，确保政策目标的科学性和可行性。其次，借助大数据与AI的动态监测能力，政府可以实时跟踪政策执行情况和减排成效，及时调整优化政策措施，确保碳减排目标的有效达成。

## （四）研究方法和创新

本文采用的研究方法包括**机器学习模型、卷积神经网络（CNN）、长短时记忆网络（LSTM）、CNN-LSTM和空间聚类分析。**CNN和LSTM是两种基础的深度学习方法。本文结合了 CNN可以用来捕捉序列中的局部特征模式的特点和LSTM可以负责处理序列中的长期依赖关系的特点，构建了CNN-LSTM模型。这样的好处是可以在处理序列前先通过卷积层压缩数据，减少LSTM需要处理的时间序列长度，从而降低计算复杂度和内存需求，提高模型训练和预测的效率。针对大规模数据集的聚类，要求算法具备良好的可扩展性，因此本文还使用了空间聚类分析法。其能够通过综合考虑对象的属性特征和空间位置，发现不同省份数据中的隐藏模式，提供有价值的信息。

## （五）文献综述

在全球气候变化的严峻挑战下，碳排放管理与减排已成为国际社会共同关注的重大议题。近年来，随着大数据与人工智能技术的迅速发展，这两项技术在环境保护、尤其是碳排放统计与减排策略中的应用日益受到学界和政策制定者的重视。

关于大数据在碳排放统计中的应用。许多研究强调了大数据在整合多源排放数据、提高监测精度方面的作用。例如，Li et al.(2020)[5]利用卫星遥感数据、地面监测站数据以及企业自报数据，构建了全国尺度的碳排放实时监测系统，显著提高了碳排放量估算的准确性。大数据技术也被应用于碳排放量的预测。Wang and Chen(2019)[6]应用机器学习算法对不同行业未来的碳排放趋势进行预测，为政策制定提供了科学依据。

关于人工智能在减排策略制定中的应用。研究显示，AI在优化能源分配、提升能效方面展现出巨大潜力。Zhao et al.(2021)[7]开发了一种基于深度强化学习的智能电网调度模型，实现了电力系统的高效运行与碳排放的显著降低。人工智能技术还能用于评估不同减排政策的潜在效果。Sun and Zhang(2020)[8]使用AI模型模拟了多种碳税政策对中国碳排放的影响，为政策选择提供了量化参考。

关于综合应用与挑战。Dong et al.(2018)[9]提出了一套结合大数据分析与AI的智慧城市碳管理框架，强调了数据驱动决策在城市规划中的作用，展示了综合技术在促进低碳城市发展中的潜力。尽管进展显著，但现有研究也指出，数据质量、隐私保护、算法透明度以及技术与政策的协同等问题仍是制约大数据与AI在碳排放管理中广泛应用的关键挑战。

综上所述，大数据与人工智能在碳排放统计与减排策略领域的应用正逐步深入，为实现精准管理、科学决策提供了新的工具和视角。然而，要充分发挥这些技术的潜力，还需克服技术、制度、伦理等方面的挑战，促进技术与政策的有效融合，为全球碳中和目标的实现贡献力量。未来的研究可进一步探索技术的具体应用场景、优化模型算法、加强跨学科合作，以及解决实施过程中遇到的实际难题。

# 二、数据处理

## （一）数据来源

本文使用的碳排放数据集整合了来自中国分部门核算和表观核算的视角，覆盖了1997年至2021年的长期时间序列，为分析中国碳排放趋势、地区差异及行业特性提供了坚实基础。数据源自权威数据库和科研团队的研究成果，确保了分析的科学性和可靠性。

本文主要进行相关数据查询的网站如下：

中国碳核算数据库：<https://www.ceads.net.cn/data/>

中国多尺度排放清单模型：<http://meicmodel.org.cn/?page_id=2345>

## （二）数据预处理和数据描述

基于对原始数据完整性的考虑，本文先使用人工筛查的方法，剔除了观测值均为0的变量。其次通过SAS统计软件对数据进行描述性统计分析，观察数据特征，合并相同类型的数据。随后，本文将每个工作表的数据存储为数组，并合并成三维数组。最后对每个样本（即每个工作表的数组）进行标准化处理，确保数据的可比性。

在中国分部门核算碳排放清单的原始数据中，通过数据预处理，本文根据工程类别一共保留了47个指标，包括：TotalEmissions（总体排放量）、Farming,Forestry,Animal Husbandry, Fishery and Water Conservancy （农业、林业、畜牧业、渔业和水利）、Coal Mining and Dressing（煤炭开采及洗选）、Petroleum and Natural Gas Extraction（ 石油和天然气开采）、Ferrous Metals Mining and Dressing（黑色金属矿开采及选矿）、Nonferrous Metals Mining and Dressing（有色金属矿开采及选矿）、Nonmetal Minerals Mining and Dressing（非金属矿开采及选矿）、Other Minerals Mining and Dressing（其他矿物开采）、Food Processing（食品加工）、Food Production（食品制造）、Beverage Production（饮料制造）、Tobacco Processing（烟草加工）、Textile Industry（纺织业）、Garments and Other Fiber Products（服装及其他纤维制品）、Leather, Furs, Down and Related Products（皮革、毛皮、羽绒及相关制品）、Timber Processing, Bamboo, Cane, Palm Fiber & Straw Products（木材加工、竹藤棕草制品）、Furniture Manufacturing（家具制造）、Papermaking and Paper Products（造纸及纸制品）、Printing and Record Medium Reproduction（印刷和记录媒介复制）、Cultural, Educational and Sports Articles（ 文化、教育及体育用品）、Petroleum Processing and Coking（石油炼制及焦化）、Raw Chemical Materials and Chemical Products（原料化学品及化学制品）、Medical and Pharmaceutical Products（医药及药品）、Chemical Fiber（化学纤维）、Rubber Products（橡胶制品）、Plastic Products（塑料制品）、Nonmetal Mineral Products（非金属矿产品）、Smelting and Pressing of Ferrous Metals（黑色金属的冶炼压制）、Smelting and Pressing of Nonferrous Metals（有色金属的冶炼压制）、Metal Products（金属制品）、Ordinary Machinery（普通机械）、Equipment for Special Purposes（特殊用途设备）、Transportation Equipment（运输设备）、Electric Equipment and Machinery（电气设备和机械）、Electronic and Telecommunications Equipment（电子和电信设备）、Instruments, Meters, Cultural and Office Machinery（仪器、仪表、文化和办公机械）、Other Manufacturing Industry（其他制造业）、Scrap and waste（废料和废物）、Production and Supply of Electric Power, Steam and Hot Water（电力、蒸汽和热水）、Production and Supply of Gas（天然气的生产和供应）、Production and Supply of Tap Water（自来水的生产和供应）、Construction（建设）、Transportation, Storage, Post and Telecommunication Services（运输、仓储、邮政和电信服务）、Wholesale, Retail Trade and Catering Services（批发、零售和餐饮服务）、Others（其他排放）、Urban（城市排放）、Rural（农村排放）。而对于每一年的排放量，本文保留了22个指标，以描述具体类型的碳排放，包括：Raw\_Coal（原煤）、CleanedCoal（清洁煤）、Other\_Washed\_Coal（其他洗煤）、Briquettes（煤球或煤块）、Coke（焦炭）、Coke\_Oven\_Gas（焦炉煤气）、Other\_Gas（其他气体燃料）、Other\_Coking\_Products（其他焦化产品）、Crude\_Oil（原油）、Gasoline（汽油）、Kerosene（煤油）、Diesel\_Oil（柴油）、Fuel\_Oil（燃料油）、LPG（液化石油气）、Refinery\_Gas（炼厂气）、Other\_Petroleum\_Products（其他石油制品）、Natural\_Gas（天然气）、Scope\_2\_Heat（范围2热能）、Scope\_2\_Electricity（范围2电力）、Other\_Energy（其他能源）、Process（工艺过程）、Scope\_1\_Total（范围1总排放）。

在中国表观碳排放清单的原始数据中，从1997年至2021年分年度一共记录了29个指标，本文保留了23个指标，包括：Raw coal total（原煤总量）、Raw Coal Indigenous production（原煤本土生产量）、Raw Coal Import（原煤进口量）、Raw Coal Export(-)（原煤出口量）、Raw Coal Stock decrease（原煤库存减少量）、Raw Coal Loss:（原煤损耗量）、Raw Coal Non-energy use（原煤非能源用途量）、Crude oil total（原油总量）、Crude Oil Indigenous production / Import / Export (-)（含义与原煤相同，只是针对原油）、Crude Oil Stock decrease / Loss / Non-energy use（对应原油的库存减少、损耗和非能源用途）、Natural gas total（天然气总量）、Natural Gas Indigenous production / Import / Export (-)（含义与原煤相同，只是针对天然气）、Natural Gas Stock decrease / Loss / Non-energy use（对应天然气库存变化、损耗和非燃烧用途）、Cement（水泥生产过程中产生的排放）、Total apparent CO2 emissions（总表观二氧化碳排放量）。

在数据预处理之后，本文选取了三个重要节点，分别为1997年、2010年和2021年，绘制了排放指标的占比图。由于指标数量过多，本文在绘图的过程中保留了排放量最多的几个指标，将其余指标相加作为一类。

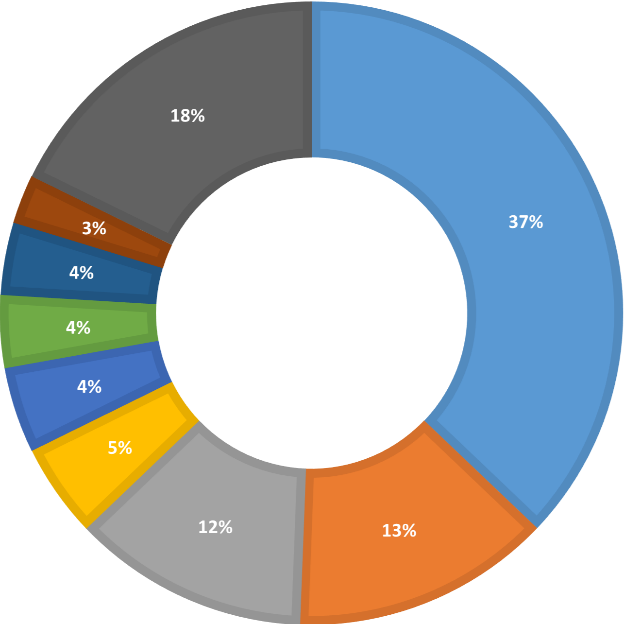
 

图 2 中国碳排放指标图例 图 3 1997年中国碳排放指标占比

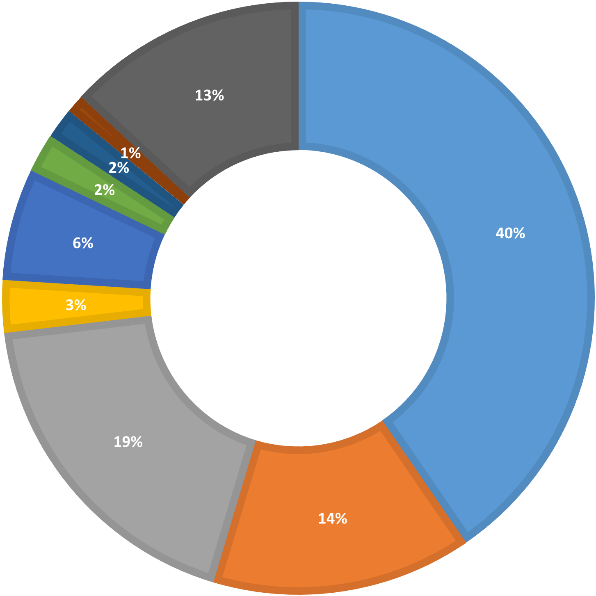
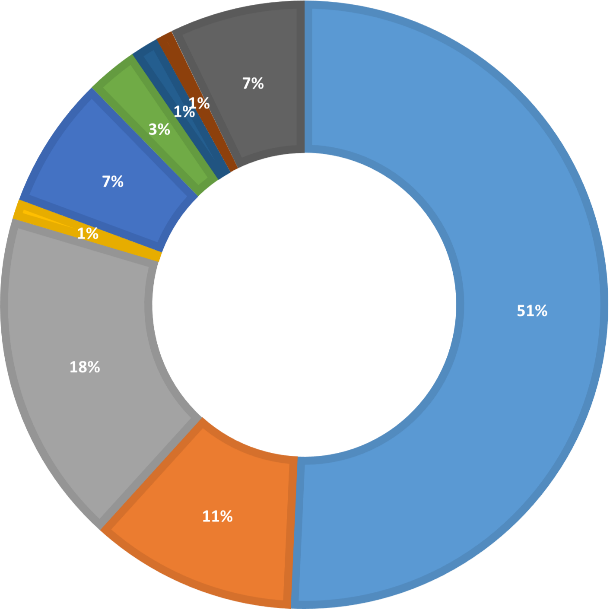
 

图 4 2010年中国碳排放指标占比 图 5 2021年中国碳排放指标占比

从图中可以看出，生产和供应电力、蒸气和热水占比最大，逐渐从37%增长到50%，符合生活常识。发电方法主要以火电为主，供热方法以煤炭为主。观察图表的变化情况，可以发现“生产和供应电力、蒸气和热水”部分占比随年份而增长，其他部分占比均有所下降。可见，煤炭是我国能源市场最主要的部分，同时煤炭在三大主要能源中产生了最多的污染。

# 三、模型构建和分析

## （一）模型假设

本文根据数据特点和研究需求，作出如下假设：

（1）假设排放数据随时间呈现一定周期性和趋势性，无突变异性大波动。

（2）假设所选模型能够精确捕获数据间的关系，预测未来趋势。

（3）假设通过标准化处理的数据的统计特性随时间变化较小。

（3）假设未来无极端外部冲击对排放量造成影响，如政策、技术革命性变化等。

（5）对于LSTM模型，假设经过处理后的输入特征之间没有多重共线性问题。

（6）对于CNN模型，假设输入数据中的每个特征都具有重要性，并且卷积核能够捕捉到这些特征的重要性和相关性。

## （二）对碳排放数据的地域考察——空间聚类分析

### 1.模型的引入

本文绘制了1997年和2021年全国各省二氧化碳排放的可视化图表，从图中可以看出，不同省份的二氧化碳排放量有显著差别，因此本文希望引入空间聚类分析模型探讨不同省份碳排放量之间的异同。

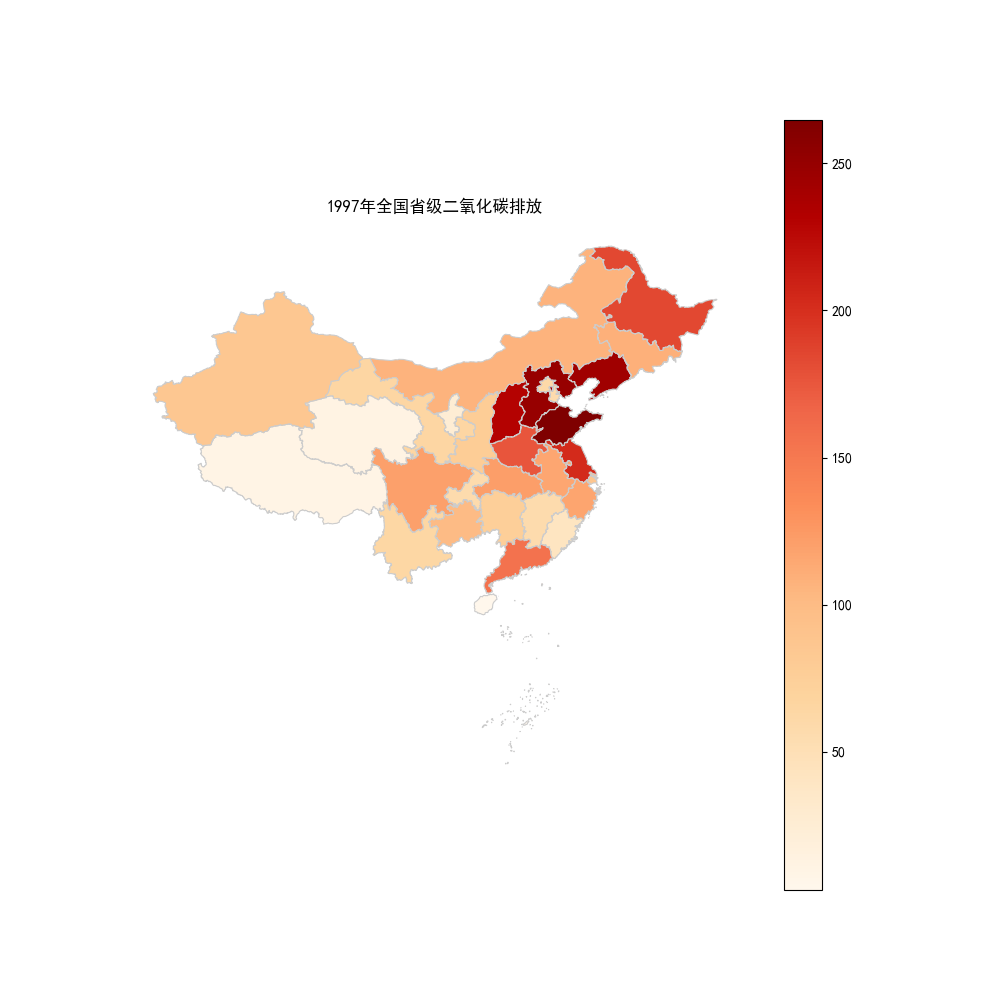
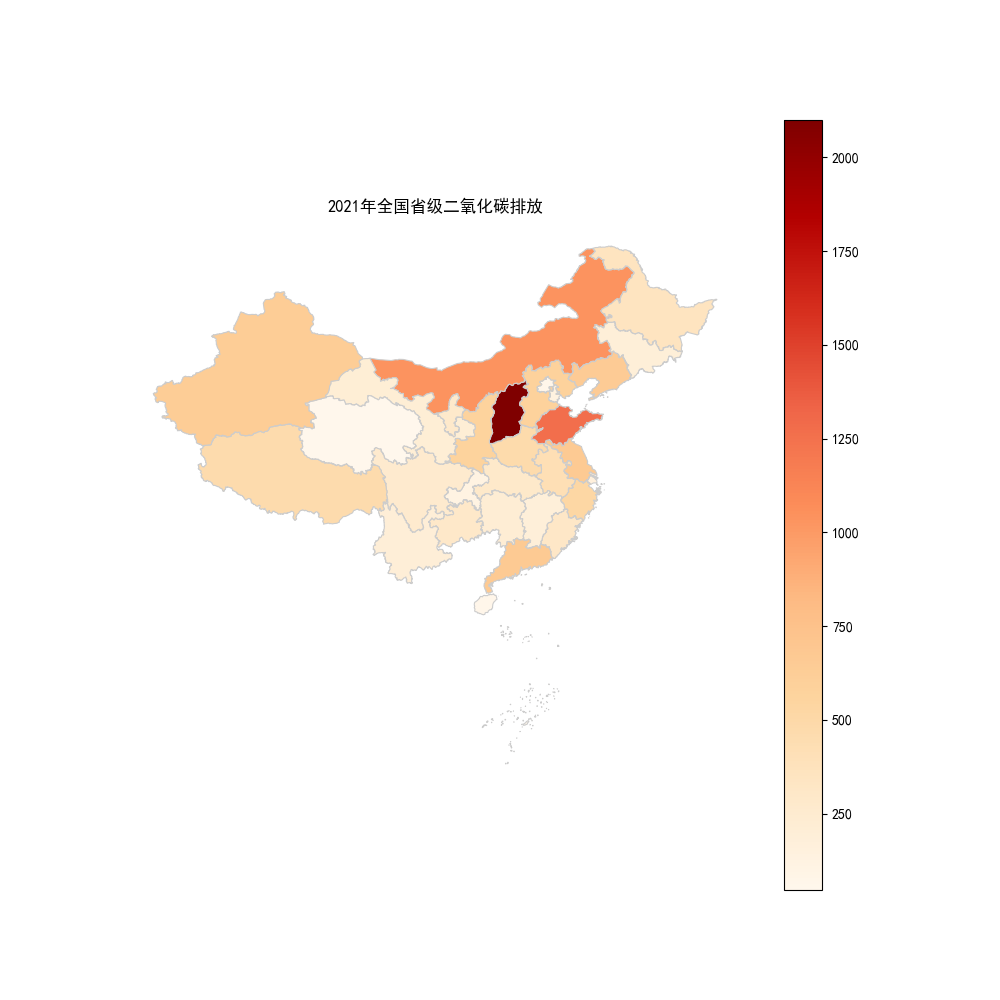
 

图 6 1997年和2021年全国各省二氧化碳排放量

### 2.空间权重矩阵

本文使用全国各个省份的碳排放数据构建空间权重矩阵，变量包括原煤、原油、天然气和水泥的碳排放量，以及总碳排放量。在先前的数据处理过程中，我们已经对数据进行了标准化处理，使所有特征具有相同尺度，这对于K-Means聚类算法至关重要。。

首先，本文通过计算距离矩阵的倒数得到空间权重矩阵，这里加入了一个极小的正值（1e-10）以避免除以零的错误。距离越小，空间权重越大，意味着地理位置相近的省份在聚类时会被认为更加相似。随后，将这个权重矩阵转换为稀疏矩阵形式，以优化计算效率。之后，本文利用构建好的空间权重矩阵与标准化后的碳排放特征向量进行点乘运算，生成了新的加权特征向量。这一步骤确保了地理位置邻近省份的碳排放特征在聚类过程中获得更高的相似度评分。

### 3.K-Means聚类

本文先确定聚类数量为5，然后应用K-Means算法对空间加权后的碳排放特征进行聚类。在程序中，本文将参数random\_state设置为42，保证每次运行结果的可复现性。

根据聚类结果，31个省份被分到了五个不同的簇，本文根据每个簇中省份的特点，对类别进行命名。分类结果见下表：

表1 空间聚类分析模型的聚类结果

|  |  |  |
| --- | --- | --- |
| 序号 | 类别 | 省份 |
| 0 | 资源型省份 | 青海、甘肃、陕西、西藏、新疆、天津、黑龙江、内蒙古、宁夏、吉林、海南 |
| 1 | 经济发达省份 | 北京、江苏、浙江、河北、广东、四川、上海 |
| 2 | 特定资源省份 | 山西 |
| 3 | 工业与能源生产省份 | 山东、辽宁 |
| 4 | 中南部综合型省份 | 湖南、湖北、安徽、江西、云南、广西、贵州、河南、福建 |

根据聚类结果，本文总结了每个簇的特点如下：

（1）序号为0的资源型省份。这个聚类包括了资源丰富的省份，特别是煤炭资源丰富的省份，如青海、甘肃、陕西、西藏、新疆、天津、黑龙江、内蒙古、宁夏、以及少量石油的海南。这些地区的碳排放显示了高碳排放与自然资源（尤其是煤炭）紧密相关的特征。

（2）序号为1的经济发达省份。此聚类包含经济发展水平较高的东部沿海省份，如北京、江苏、浙江、河北、广东、四川、上海，它们在工业化和城市化进程中碳排放量较大，经济活动活跃，同时具有较高的能源消费量，包括煤炭、石油和天然气。这些地区的碳排放体现了经济发展水平与碳排放量的正相关，表示它们面临着更紧迫的节能减排压力。

（3）序号为2的特定资源省份。此聚类仅包含山西省，山西省是我国煤炭生产的主要基地，碳排放特征显著不同。其单独成簇表明了山西作为煤炭大省的独特地位。

（4）序号为3的工业与能源生产省份。此聚类包括山东省和辽宁省，这两个省份有较高的重工业基础，同时辽宁还有一定的石油资源。它们的碳排放特征可能与重工业生产和能源消费模式相关联，揭示了工业活动对碳排放的显著影响。

（5）序号为4的中南部综合型省份。此聚类主要包含中南部省份，如湖南、湖北、安徽、江西、云南、广西、贵州、河南、福建。这些省份的碳排放特点与农业、制造业和部分能源生产相关，尤其是湖南、湖北等省份的农业和制造业较为发达，而云南、贵州则可能因水电资源丰富而有独特的能源结构。这些地区的碳排放体现了多样化的经济结构和能源消费模式。

## （三）对碳排放数据的拟合——机器学习和深度学习模型

### 1. 机器学习模型

本文选择三个机器学习模型对排放数据进行分析和拟合，包括线性回归模型、随机森林回归模型和梯度提升回归模型。线性回归模型主要用于探索排放与特征线性关系；随机森林回归模型可以处理非线性或多因素交互关系；而梯度提升回归模型能够挖掘更深层次的关系，提高拟合能力。

随后，本文对模型进行评估，得到的结果如下表所示：

表2 机器学习模型性能评估

|  |  |  |
| --- | --- | --- |
| 模型 | 均方误差(MSE) | 决定系数 |
| 线性回归 | 176929.71024 | 0.97391 |
| 随机森林回归 | 127452.55651 | 0.98120 |
| 梯度提升回归 | 116332.43806 | 0.98284 |

均方误差（MSE）表示拟合值与实际值之间的平均误差的平方。因此，MSE越小，表示模型的拟合能力更接近真实值。根据这一指标，梯度提升回归模型表现最好。

表示模型对目标变量的解释能力，取值范围在 0 到 1 之间。越接近 1，说明模型能够很好地解释目标变量的变异。根据这一指标，梯度提升回归模型同样表现最好。

根据梯度提升回归模型在MSE和上的优异表现，我们可以认为它在拟合中国碳排放量方面具有较高的准确性和解释能力，随机森林模型次之，而线性回归模型在这两个指标上稍逊一筹。

### 2. 深度学习模型

本文选择三个深度学习模型对排放数据进行分析和拟合，包括CNN模型、LSTM模型和CNN-LSTM模型。

CNN模型的模型结构分为四层，包括输入层、卷积层、最大池化层和全连接层。我们在输入层输入维度为各个碳排放指标的数据。卷积层包含2个卷积层，分别有32和64个过滤器，卷积核大小为3，能够输出过滤后的数据。最大池化层能对卷积层的输出进行下采样。最后再由全连接层将卷积层的输出映射到最终的拟合值。

LSTM模型的模型结构分为三层，包括输入层、LSTM层和全连接层。输入层和全连接层的作用与CNN模型相同，LSTM层包含两个隐藏层，每层128个神经元，能够输出数据。

考虑到两种模型各有优势，因此本文结合LSTM模型和CNN模型的特性，构建了一个混合模型。先用CNN捕捉局部特征，再通过LSTM捕获长期依赖关系，以达到更优的拟合效果。CNN-LSTM模型的模型结构分为五层，包括输入层、一维卷积层、重塑层、长短期记忆层和全连接层。输入层和全连接层的作用与前两个模型相同。一维卷积层主要利用了CNN模型的特性，能够捕捉局部特征。它能通过卷积核在输入序列上移动，对输入数据应用不同的滤波器来检测不同特征。它还会使用ReLU激活函数增强非线性表达能力，并且应用最大池化来降低空间维度，减少计算量。重塑层是CNN到LSTM的过渡，它将数据从卷积后的形状转换为适合LSTM处理的序列形式，使得序列长度成为LSTM处理的第一维度。长短期记忆层主要利用了LSTM模型的特性，它能够控制信息的流入、存储和流出，有效处理时间序列中的时间依赖关系。

随后，我们对模型进行评估，使用测试集计算模型的均方误差（MSE）、平均绝对误差（MAE）和决定系数（）。结果如下表所示：

表3 深度学习模型性能评估

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 均方误差(MSE) | 平均绝对误差(MAE) | 决定系数（） |
| CNN | 0.00688 | 0.03314 | 0.95107 |
| LSTM | 0.01065 | 0.03113 | 0.97240 |
| LSTM-CNN | 0.00842 | 0.03198 | 0.98945 |

在均方误差上，CNN模型的MSE为0.00688，为三个模型中最低，说明其在拟合误差的平方和方面最为精确。LSTM-CNN混合模型的MSE为0.00842，优于LSTM模型的0.01065，这表明混合模型在减小拟合值与实际值之间平方误差的累积上取得了更好的平衡。

在平均绝对误差上，LSTM和LSTM-CNN混合模型的MAE非常接近，分别为0.03113和0.03198，这说明两者在拟合值与实际值的平均偏差上表现相似，且均优于CNN模型的0.03314。这表明这两种模型在提供接近实际值的拟合方面做得更好。

在决定系数上，LSTM-CNN混合模型的值显著提升至0.98945，远高于LSTM模型的0.97240和CNN模型的0.95107。这意味着混合模型不仅在减少拟合误差方面表现优秀，而且在解释数据变异和趋势拟合方面达到了最佳状态，模型的拟合值与实际观测值高度相关，几乎解释了数据变异的98.95%。

综上分析，LSTM-CNN混合模型不仅在MSE和MAE上达到了平衡，而且在决定系数上有了显著提升，这表明该模型在拟合碳排放数据方面综合性能最佳。它成功地结合了CNN对局部特征的有效提取和LSTM对长期依赖关系的良好捕捉，达到了高水平的拟合准确性和趋势拟合能力。

## （四）对碳排放数据的预测——ARIMA模型

将数据转化为时间序列后进行平稳性检验（ADF）检验，发现数据不平稳，进行两次差分后得到p-value=3.18e-09：

表4 碳排放数据差分结果

|  |  |
| --- | --- |
| 差分次数 | p-value |
| 0 | 0.68203 |
| 1 | 0.07465 |
| 2 | 3.17730e-09 |

可以认为此时数据通过了平稳性检验，数据平稳。观察差分后数据的ACF、PACF图：

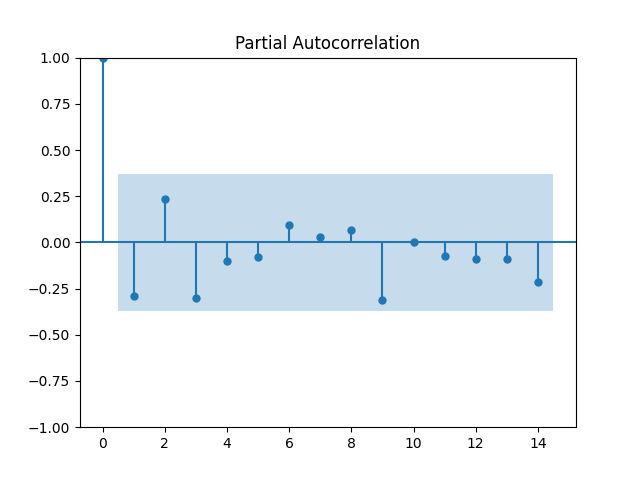
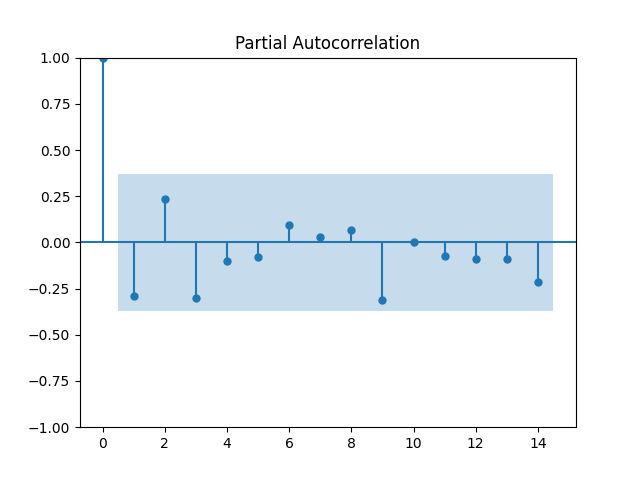


图 7 差分后数据的ACF、PACF图

可以发现图像截尾，定阶AR=1，MA=1，那么ARMIA模型为ARIMA（1,2,1）。同时使用最优子集法进行定阶，发现AIC、BIC指标稍高于ARIMA（1,2,1）：

表5 ARIMA参数表

|  |  |  |  |
| --- | --- | --- | --- |
| （p，d，q） | Log Likelihood | AIC | BIC |
| （1，2，1） | -187.233 | 380.465 | 384.462 |
| （0，2，1） | -190.009 | 384.018 | 386.682 |
| （1，2，0） | -189.697 | 383.394 | 386.058 |

故使用该模型进行预测。

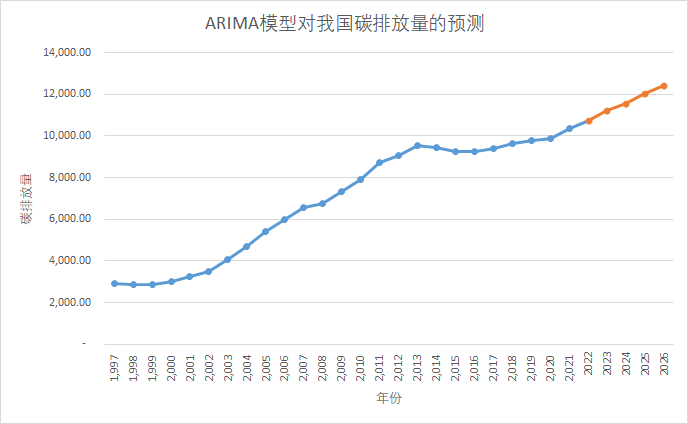


图 8 ARIMA模型对我国碳排放量的预测

可以看出基于ARIMA模型的时间序列预测值逐年递增。

# 四、结论和建议

## （一）研究结论

本文基于1997—2021年中国分部门核算碳排放清单数据以及中国表观碳排放清单数据，在大数据与人工智能视角下，考察不同地域的碳排放数据，并预测未来两年的碳排放趋势，得出如下结论：①在全国整体层面：中国碳排放总量在此期间显著增长，反映了工业化与经济增长带来的环境压力，同时也彰显出控制碳排放、推动绿色低碳发展的紧迫性。通过大数据与人工智能技术的综合应用，研究揭示了排放趋势的周期性与增长模式，为全国性的碳排放控制提供了数据支持与趋势洞察；②在分部门碳排放统计层面，人工智能技术通过优化能源利用效率在减排中发挥了核心作用。研究详细分析了各行业（如能源生产、工业、交通、农业等）的排放特征，发现人工智能技术如深度强化学习在智能电网调度、工艺流程优化、交通系统智能化等方面的应用，有效提升了能效，降低了碳排放。这表明，通过技术革新推动产业升级，特别是提高能源生产和消耗效率，是实现部门减排的关键路径；③在分省份碳排放统计层面，利用人工智能技术对省份数据进行空间聚类分析，识别了排放密集区域和减排潜力较大的板块，为地方制定针对性的减排策略提供了科学依据。

## （二）减排策略

在对中国分部门核算碳排放清单（1997-2021年）的分析基础上，我们发现碳排放总量从1997年的约2923.86百万吨二氧化碳当量增长至2021年的超过10356.3百万吨二氧化碳当量，显示出长期增长趋势。针对这一情况，我们推测，在较长一段时间内，全国二氧化碳排放将在长期体现出增长态势。结合各分部门的具体排放特征，可以提出针对性的减排办法：

（1）能源生产与供应部门：对于煤炭行业，应重点推进煤炭清洁高效利用技术，如超临界、超超临界燃煤发电技术，以及碳捕获技术。同时，逐步降低煤炭在能源结构中的比例，增加天然气等低碳能源占比。对于石油和天然气行业，应提高石油和天然气的勘探与开采效率，同时发展天然气分布式能源系统，以天然气替代煤炭，减少碳排放。

对于电力行业，应大力发展可再生能源，提高其在能源供应中的比重，鼓励农村使用太阳能发电，优化电网调度策略，确保可再生能源有效接入和利用。

（2）工业部门：对于钢铁、水泥和化工行业，可以推广短流程炼钢、干熄焦技术，减少工艺过程中的碳排放。对于制造业，可以推动能效提升项目，如采用高效电机、节能灯具，优化生产流程，以及通过数字化转型提高整体能源管理效率。采用高效加热和冷却系统，以及优化供应链减少运输排放。

（3）交通部门：鼓励使用公共交通，发展电动汽车和氢能车辆，建设充电基础设施。

（4）农业与林业部门：对于农业，应减少化肥和农药使用，发展有机农业，同时利用农业废弃物进行生物质能源生产。对于林业，应扩大植树造林和森林保护面积，同时发展森林碳汇项目。

（5）政策与市场：应完善全国碳排放权交易市场，设定合理的碳价，激励企业减排。并对低碳技术、产品和服务给予税收减免或补贴，对高碳排放产品征收环境税，鼓励制造业企业使用回收或可再生原材料，减少碳足迹。同时加大低碳技术研发投入，参与国际气候治理，分享减排经验和技术。

# 参考文献

1. 赵 成 柏,毛 春 梅.基 于ARIMA和BP神 经 网 络 组合模型的我国碳排放强度预测[J].长江流域资源与环境,2012,21(6):665-671.
2. 曾晓莹,邱荣祖,林丹婷,等.中国交通碳排放及影响因素时空异质性[J].中国环境科学,2020,40(10):4304-4313.
3. 郭春丽, 易信. “双碳”目标下的中国经济增长: 影响机制, 趋势特征及对策建议[J]. 经济学家, 2022, 1(7): 24-33.
4. 钟进, 李宗航. 基于趋势性时间序列的全国碳排放量预测研究[J]. 运筹与模糊学, 2023, 13(4): 3870-3881.
5. Li, X., Wang, S., Zhang, L., & Liu, Z. (2020). Real-time monitoring and estimation of national CO2 emissions using satellite observations and machine learning techniques. Remote Sensing of Environment, 246, 111890.
6. Wang, Y., & Chen, W. (2019). Predicting carbon emissions in different sectors using machine learning: A case study of China. Applied Energy, 236, 1248-1259.
7. Zhao, Y., Wu, J., Xu, M., & Wang, J. (2021). Deep reinforcement learning-based energy management system for reducing carbon emissions in smart grids. IEEE Transactions on Smart Grid, 12(4), 3456-3467.
8. Sun, C., & Zhang, Y. (2020). Evaluating the impact of carbon tax policies in China using artificial intelligence algorithms. Energy Policy, 144, 111673.
9. Dong, H., Zhang, P., & Liu, Y. (2018). A big data and artificial intelligence enabled smart city framework for carbon management. Journal of Cleaner Production, 172, 45-56.
10. Tong, D., J. Cheng, Y. Liu, S. Yu, L. Yan, C. Hong, Y. Qin, H. Zhao, Y. Zheng, G. Geng, M. Li, F. Liu, Y. Zhang, B. Zheng, L. Clarke and Q. Zhang (2020), Dynamic projection of anthropogenic emissions in China: methodology and 2015-2050 emission pathways under a range of socio-economic, climate policy, and pollution control scenarios, Atmos. Chem. Phys., 20, 5729-5757.

# 附录

1. 词云图R代码

# 导入所需的包

library(tm)

library(wordcloud)

library(wordcloud2)

# 读取文本文件，转换成文档对象

text <- readLines("F:\\统计建模[\\Code\\R](file:///\\Code\\R)\\processed\_text.txt", warn = FALSE)

corpus <- Corpus(VectorSource(text))

# 文本预处理

corpus <- tm\_map(corpus, content\_transformer(tolower)) # 转换为小写

corpus <- tm\_map(corpus, removePunctuation) # 去除标点符号

corpus <- tm\_map(corpus, removeNumbers) # 去除数字

corpus <- tm\_map(corpus, removeWords, c(stopwords("english"), "的","和","为"))

# 创建词频矩阵

dtm <- DocumentTermMatrix(corpus

# 计算词频，创建数据框

word\_freq <- colSums(as.matrix(dtm))

freq <- data.frame(word = names(word\_freq), freq = word\_freq)

# 过滤词频低于10的词

freq <- freq[freq$freq >= 5, ]

# 生成词云图

freq\_df <- data.frame(word = character(0), freq = numeric(0))

freq\_df <- rbind(freq\_df, freq)

wordcloud2(data = freq\_df, shape = 'circle')

# 致谢

感谢李亚杰老师对论文的指导。您的悉心指导与严谨治学的态度是我们学术道路上的明灯，从论文选题到框架搭建，再到无数次的修改与完善，每一步都凝聚着您的智慧与心血。

感谢王若禹同学给予的宝贵意见。

感谢所有在本文献引用中提及的学者，你们的研究成果为我们的研究奠定了坚实的基础，激发了我们新的思考与灵感。