

清 华 大 学

# 综 合 论 文 训 练

题目：山西省农田土壤有机碳演变特征及趋势预测

系 别：未央书院

专 业：数理基础科学+能源与动力工程

姓 名：杨思颖

指导教师：赵永敢 副研究员

2024 年 6 月 17 日

## 中文摘要

土壤有机碳是全球碳汇的重要组成部分，在充分发挥土壤固碳能力和减少土壤呼吸作用方面具有重要作用，对于大气减碳具有积极的影响。在可持续发展目标下，山西省可以通过限制企业的碳排放，并加强固碳技术，例如增加土壤碳汇，以实现大气碳中和的目标。研究山西省土壤有机质的时空变化特征和未来的应用前景对于山西省农业生产和农田固碳减排具有重要意义。

本研究整理了全国二普（1982年）、山西省各县市耕地地力评价与利用文献数据（2012年）和实测数据（2022年），并结合 Google Earth Engine 和国家青藏高原数据中心的气候、地形、植被等数据。运用统计学、地统计学、皮尔逊相关分析和机器学习模型，分析山西省农田土壤有机质的时空演变规律和预测前景，为山西省有机质提升和农田固碳减排提供科学依据。研究结果如下：

(1) 通过描述性统计发现，自 1982 年至 2022 年，山西省耕地土壤有机质含量逐渐增加。地统计学方法计算的半方差函数显示，1982 年山西省不同地区的有机质呈现中等强度的空间相关性，而 2012 年和 2022 年的有机质含量呈现出强空间相关性。克里格插值绘图显示，1982 年山西省耕地有机质含量由西向东递增，而 2012 年和 2022 年呈西北向东南递增的趋势。

(2) 通过皮尔逊相关分析发现，有机质含量与 pH 值呈较弱的线性负相关，与全氮和碱解氮有强线性正相关；与地形、气候和植被因子之间的线性关系不明显。有机质含量与人为肥料输入之间存在很强的相关性。

(3) 通过建立线性回归、支持向量机、随机森林、XGBoost 和 LightGBM 模型，研究了以上因素与有机质含量之间的关系，并比较了不同模型的预测效果。结果显示，在该预测情景下，LightGBM 和 XGBoost 模型表现最佳。本研究设定了未来 10 年各指标的变化情景，并利用这两个模型预测了该情景下山西省农田土壤的有机质含量分别为 20.0g/kg 和 19.6g/kg。

本研究通过描述现象、明确影响因素和预测未来三个步骤，深入了解山西省农田土壤有机质的时空变化情况及其作用机制，并对未来有机质含量的变化进行了可行的预测，期望对于山西省耕地管理和固碳措施的制定提供有益的参考。

**关键词：**山西省耕地；土壤有机质；地统计学；相关性分析；机器学习模型

## ABSTRACT

Soil organic carbon is an important component of the global carbon sink and plays a crucial role in maximizing soil carbon sequestration capacity, thereby impacting carbon emission reduction. Understanding the spatiotemporal variations and future prospects of soil organic matter in Shanxi Province is significant for agricultural production and carbon sequestration.

Statistical analysis, geostatistics, Pearson correlation analysis, and machine learning models were employed to analyze the spatiotemporal patterns and predictive prospects of soil organic matter in farmland in Shanxi Province.

(1) Descriptive statistics revealed a gradual increase in soil organic matter content in farmland of Shanxi Province from 1982 to 2022. Geostatistical analysis showed moderate spatial correlation of organic matter in different regions of Shanxi Province in 1982, while strong spatial correlation was observed in 2012 and 2022. Kriging interpolation maps displayed an increasing trend of organic matter content from west to east in 1982, and from northwest to southeast in 2012 and 2022.

(2) Pearson correlation analysis indicated a weak linear negative correlation between organic matter content and pH value, while a strong linear positive correlation was observed with total nitrogen and alkaline hydrolyzable nitrogen. There was a strong correlation between organic matter content and anthropogenic fertilizer input.

(3) By establishing linear regression, support vector machine, random forest, XGBoost, and LightGBM models, the relationships between the aforementioned factors SOM were studied. LightGBM and XGBoost models performed the best. These two models were used to predict SOM in farmland soil in Shanxi Province under certain scenarios, yielding predicted values of 20.0g/kg and 19.6g/kg, respectively.

This study provides in-depth insights into the mechanisms of SOC in farmland in Shanxi Province, along with feasible predictions of future organic matter content.

**Keywords:** Shanxi Province farmland; soil organic matter; geostatistics; correlation analysis; machine learning

# 目 录

<b>第1章 引言</b> .....	<b>1</b>
1.1 研究背景及意义 .....	1
1.2 山西省土壤有机质研究进展 .....	3
1.2.1 土壤有机质时空演变特征研究 .....	3
1.2.2 土壤有机质空间变异性研究 .....	3
1.2.3 土壤有机质时空演变的影响因素 .....	4
1.3 土壤有机质预测研究进展 .....	5
1.4 研究内容 .....	6
1.5 技术路线 .....	7
<b>第2章 数据来源与分析方法</b> .....	<b>9</b>
2.1 研究区概况 .....	9
2.2 数据来源 .....	10
2.2.1 土壤数据来源 .....	10
2.2.2 气候、地形、植被数据来源 .....	11
2.2.3 农资数据来源 .....	12
2.3 研究方法 .....	12
2.3.1 地统计学理论 .....	12
2.3.2 相关性分析 .....	14
2.3.3 多元线性回归 .....	15
2.3.4 支持向量机 .....	15
2.3.5 随机森林 .....	15
2.3.6 XGBoost .....	16
2.3.7 LightGBM .....	16
<b>第3章 山西省农田土壤有机质时空演变特征</b> .....	<b>17</b>
3.1 山西省农田土壤有机质时空变化特征 .....	17
3.2 山西省农田土壤养分时空变化特征 .....	17
3.3 山西省耕地土壤肥力指标的时空分布特征 .....	20

3.3.1 耕地有机质的时空空间分布情况 .....	20
3.3.2 耕地全氮的时空空间分布情况 .....	27
3.4 本章小结 .....	29
<b>第 4 章 山西省农田土壤有机质影响因素分析 .....</b>	<b>31</b>
4.1 土壤有机质与养分的关系 .....	31
4.2 土壤有机质与地形、气候、植被的关系 .....	33
4.2.1 各影响因素的时空分布 .....	33
4.2.2 地形、气候、植被因子对有机质的影响 .....	37
4.3 土壤有机质与农资投入的关系 .....	39
4.4 本章小结 .....	41
<b>第 5 章 山西省农田土壤有机质变化趋势预测 .....</b>	<b>42</b>
5.1 输入模型的因子准备 .....	42
5.2 模型构建 .....	43
5.2.1 多元线性回归 .....	43
5.2.2 支持向量机回归 .....	46
5.2.3 随机森林 .....	47
5.2.4 XGBoost .....	48
5.2.5 LightGBM .....	50
5.3 模型效果比较 .....	51
5.4 设定未来场景预测 .....	52
5.5 本章小结 .....	55
<b>第 6 章 结论与展望 .....</b>	<b>56</b>
6.1 结论 .....	56
6.2 不足与展望 .....	57
<b>插图索引 .....</b>	<b>58</b>
<b>表格索引 .....</b>	<b>60</b>
<b>参考文献 .....</b>	<b>62</b>
<b>致    谢 .....</b>	<b>66</b>
<b>声    明 .....</b>	<b>67</b>

附录 A 外文资料的书面翻译 .....	68
----------------------	----

# 第1章 引言

## 1.1 研究背景及意义

近几十年来，全球范围的工业化和城市化消耗了大量的非可再生能源，排放了巨量温室气体，导致地球气温不断升高，带来了一系列不可逆的环境问题<sup>[1]</sup>。自 1850 年至 2022 年，地球大气中的平均二氧化碳浓度从 285ppm 急剧增加到了 418ppm<sup>[2]</sup>，大气二氧化碳浓度的增速也在不断升高。习近平总书记在第七十五届联合国大会上承诺“我国二氧化碳排放力争于 2030 年前达到峰值，努力争取 2060 年前实现碳中和”，以此积极响应联合国政策。支撑碳中和的技术涉及到各行各业，为达成碳中和目标，我国的首要任务应是深度脱碳，此外也可以使用负排放和应用碳汇为能源系统增加灵活性<sup>[3]</sup>。

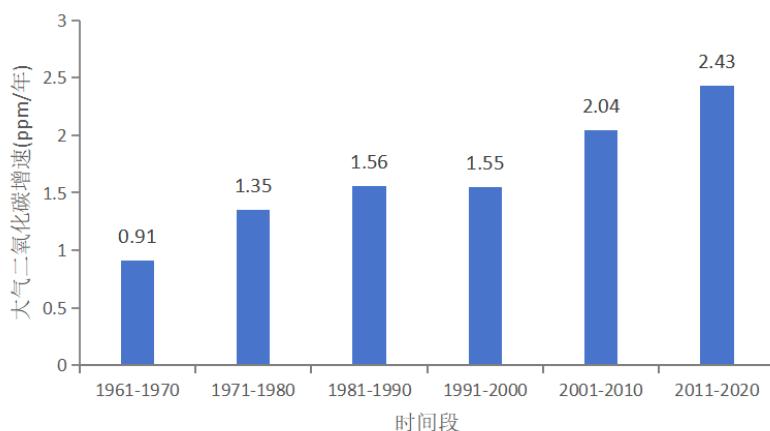


图 1.1 大气二氧化碳增速<sup>[4]</sup>

碳汇是减少大气中二氧化碳浓度的有效途径，不仅能够很好地辅助实现双碳目标，而且还可为能源系统的发展提供空间。全球共有五个主要的碳汇，它们互相链接，碳在其之间循环。其中土壤碳汇包含有机碳汇和无机碳汇，但土壤的有机碳汇可达无机碳汇近 2 倍。土壤有机碳 (SOC) 存在于土壤有机质 (SOM) 中，占 SOM 约 58%<sup>[5]</sup>。土壤有机质中含有的有机碳 (OC) 比全球植被和大气中的总和还要多，其中一小部分的碳释放并转化为二氧化碳或甲烷时，也会导致这些温室气体的大气浓度的相应的变化，对碳的吸收也同理<sup>[6]</sup>。因此，我们应该通过各

种措施减少土壤有机碳的矿化和呼吸作用，增加有机碳的库存容量，为温室气体减排和土壤质量提升起重要作用。

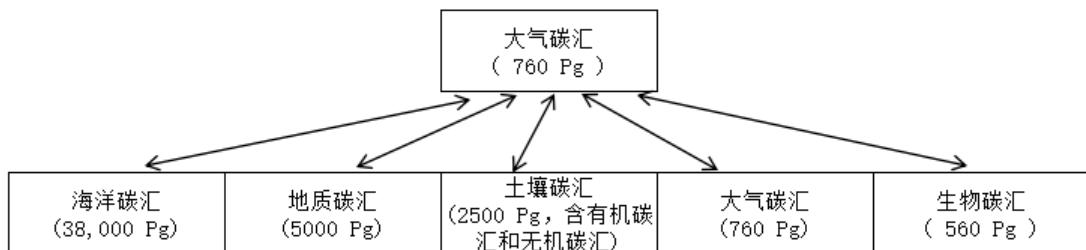


图 1.2 全球碳汇<sup>[7]</sup>

据测算，全球 2m 深土壤中储存的 SOC 达到了 24000 亿吨，是全球矿质燃料燃烧排放二氧化碳当量（89 亿吨）的 270 倍。可见，SOC 对土壤固碳、负碳排放具有积极的意义。法国农业部部长在第 21 届联合国气候变化大会期间提出了“千分之四计划”，他认为全球 2m 深土壤的有机碳储量只要每年增加 4%，即可抵消当年全球矿质燃料燃烧的碳排放<sup>[8]</sup>。相关研究表明，我国 SOC 固碳潜力为 105~198 Tg C yr<sup>-1</sup>，无机碳为 7~138 Tg C yr<sup>-1</sup>，到 2050 年累计固碳量达到 11Pg，可抵消每年工业碳排放的 20%<sup>[9]</sup>。因此，探究我国土壤有机质的现状和未来十分具有现实意义。

在农业生产中，土壤有机碳直接影响了土壤的保肥性<sup>[10]</sup>，是维持农田高产稳产的基础。提高农田有机碳含量对于提升土壤肥力、保护耕地具有重大意义，可服务我国“严防死守 18 亿亩耕地红线”的目标，为夯实大国粮仓“耕基”提供基础。然而，农田土壤有机碳含量直接受到人类活动的干预，是高效可控的，研究农田土壤有机碳可对未来提出切实可行的建议。

山西省是我国的煤炭大省，煤炭资源开发强度大，高耗能产业为主，其经济发展与煤炭消费联系非常紧密，这也造成了二氧化碳排放量过高的现状。据估算，2018 年山西省二氧化碳排放量高达 8.55 亿吨，位列全国第二，人均排放量 23 吨 / 人，也远超其他省份。山西省急需促进低碳发展，承担起 3060 双碳目标的责任。为保证山西经济发展，一味的关停碳排放企业（碳源）会造成巨大的经济损失，需要强化固碳技术（碳汇）才能高效的助力双碳目标。山西省耕地面积 7205 万亩，是一个潜在的碳库，增加农田土壤有机碳含量、增强土壤碳汇效应能够有效的为行业发展创造空间。

为帮助山西省耕地发挥碳汇作用、助力山西省达成双碳目标，本研究以山西

省农田土壤有机质为研究对象，探究其 1982 年至 2022 年的时空演变，结合多源数据（气候、地形、植被等），搭建一系列模型，了解土壤有机质的影响因素，预测其未来状况。由此期望为山西省农田土壤肥力管理、固碳管理提供数据支撑，也对未来的多源数据变化下的土壤有机质含量变化提供可行的预测和应对指引，充分发挥农田土壤碳汇作用，助力双碳目标达成。值得注意的是，有机质含量的 0.58 倍即为土壤中的有机碳含量，因此有机质含量可以作为评估土壤有机碳水平的指标。在研究和数据记录中通常记载有机质数据，因此后文均以有机质含量进行分析。

## 1.2 山西省土壤有机质研究进展

本部分内容聚焦于山西省，综述了前人对山西省土壤有机质变化的相关研究报道，并重点介绍了山西省农田土壤有机质的研究，同时简要涵盖了林地和草地等其他类型。

### 1.2.1 土壤有机质时空演变特征研究

一些研究者主要进行了现象对比的研究。例如李铮<sup>[11]</sup>利用第二次土壤普查资料，并按照全国统一的土壤有机质分级标准和山西省的生产实践，对山西省耕地土壤的有机质分区域进行了评述。朱静的研究<sup>[12]</sup>表明，如皋市的土壤有机质含量在近 20 年中没有特别的空间分布上的变化，而时间上一直呈上升趋势。解文艳<sup>[13]</sup>在山西省设置了 75 个监测点，对 2007 年和 2017 年的土壤肥力数据进行了分析，发现有机质含量增加、全氮和速效钾含量大幅增加，但是不同地区存在差异，例如晋中晋南有机质增加，而其他地区下降。

一部分研究者在进行现象描述的同时，会结合分级指标进行分级，评估土壤水平。例如胡克林<sup>[14]</sup>对北京大兴区的研究得出，1980 年到 2000 年，有机质含量表现为由低向高逐级累积递推的规律。于婧文<sup>[15]</sup>对山西省寿阳县 1998 年到 2008 年的土壤有机质按照国家养分分级指标进行了分级，发现 2008 年有机质分级主要为四级，而 1998 年指标大多为三级。

### 1.2.2 土壤有机质空间变异性研究

前人关于土壤有机质空间变异性研究大致分为两类，第一类为较为常见的经典统计学分析方法，即通过计算平均值、最大最小值、极差、标准差和变异系数，得到土壤性质的变异性特征。例如林小丁<sup>[16]</sup>采用描述性统计方法研究了陕西

省土壤有机质含量，结果显示 1980 年至 2020 年，土壤有机质含量在 4 个阶段的中值依次为 11.0、11.1、13.6 和 18.5 g/kg，整体呈显著上升趋势。张建杰<sup>[17]</sup>发现山西省临汾盆地土壤有机质的变异系数为 46.10%，属于中等变异程度。然而，传统的统计指标只能反映土壤属性数据的统计特征或分布趋势，缺乏对空间分布的描述。

地统计学能够弥补这一不足，因而也成为了当前的研究热点和主流。采用半方差函数来量化区域化变量的空间变异性特征，并获取参数，如块金值，能够解释区域化变量空间变异性的强弱，从而弥补了经典统计学中忽略空间方位的不足。在半方差函数的基础上，可以进行克里格插值，填补预测空间中缺失数据部分的土壤有机质含量。宋莎<sup>[18]</sup>的研究涉及四川省双流县土壤有机质的空间变异，发现该县土壤有机质具有强烈的空间自相关性。申若禹<sup>[19]</sup>对山西省耕地的研究发现土壤类型是土壤有机质含量空间变异的重要影响因素。王国芳<sup>[20]</sup>对山西省永济市试验农田进行了基于半变异函数和空间自相关理论的分析，研究了不同深度土层有机质含量的空间相关性和聚集特征，并采用克里格插值方法对各土层的有机质含量进行了预测。

前人研究对于土壤有机质的时间和空间变化已经取得了一定的研究成果。这些研究从宏观的角度对山西省土壤有机质含量的变化进行了现象描述，并对山西省的耕地质量有了深入的认识。然而，在山西省有机质的时间变化和空间变化的现象描述方面还存在一些不足之处。大多数研究仅针对某一年的数据进行了空间分布特征的分析，缺乏对较长时间跨度的考察。此外，研究范围主要集中在单个区县，区域尺度较为有限，对山西省在时间和空间两个尺度上有机质变化的认知和掌握仍然需要进一步明确。

### 1.2.3 土壤有机质时空演变的影响因素

众多研究认为土壤性质的空间变异主要受到系统变异和随机变异的影响。系统变异多指自然因素，比如成土母质、气候、水文、地形、地貌、生物等的差异。随机因素多指人为因素，如施肥、灌溉、种植模式等人为耕作管理措施。

在研究影响因素时，通常会进行定性研究和定量研究。定性研究结合历史事实的时空情况与有机质演变的时空比对，例如将种植模式的改动时间与土壤有机质含量的变化时间进行比对。朱静<sup>[21]</sup>通过比对土壤有机质的时空演变和秸秆还田及土壤管理的实施时间，认为时空演变与秸秆还田面积减少、农业产业结构调整以及土壤质地等因素相关。胡克林<sup>[22]</sup>研究表明，秸秆还田和施用有机肥是北京大

兴区农田土壤有机质含量普遍上升的原因，随着作物产量的不断提高，部分地区的有机质入不敷出，呈下降的趋势。

定量研究则采用各种定量方法，确定各因子对土壤有机质时空变化的影响程度。李浩<sup>[23]</sup>通过观察土壤有机质回归方程的系数发现乡镇区划在解释土壤有机质的空间分异方面具有最强的能力，占据了 38.7% 的变异解释量。其次，土壤类型（20.9%）和高程（20.7%）也对土壤有机质的空间变异起到显著的解释作用。赵明松<sup>[24]</sup>研究发现土壤质地对有机质含量变异的影响比年均温更大，二者解释变异的比例分别为 32.0% 以及 23.4%。

定性研究通过时间和空间的重叠，理解了农业政策和耕作习惯对土壤有机质含量的影响。然而，这些研究不能产生定量结论，且其说服力较弱。相比之下，定量研究方法可以通过量化的数据来理解各种影响因素的影响程度，甚至比较不同因素之间的影响强度差异。通过定量结论，人们可以深入了解如何通过改变环境变量、耕作方式等影响因素，进一步优化土壤有机质含量。然而，针对山西省土壤有机质含量影响因素的定量研究还相对较少，未来的研究需要结合多源数据，加强定量比较分析。

### 1.3 土壤有机质预测研究进展

对有机质未来的预测是近年来的研究热点。在较为基础的研究中，研究者使用了多元线性回归模型来进行预测。例如，李浩<sup>[25]</sup>设置了增产、平收、减产的不同情境，使用多元线性回归法对秭归县有机质 2019-2013 年的时空分布进行了预测。然而，实际证明环境因子和土壤养分含量之间的关系不能完全用线性来概括。因此，研究者开始利用非线性模型进行预测，例如支持向量机（SVM）、BP 神经网络和随机森林（RF）等。黄婷<sup>[26]</sup>通过比较支持向量机模型分类结果与 BP 神经网络模型、判别法和聚类分析法的分类结果，发现支持向量机模型在土壤基础肥力评价方面的结果更可靠。该研究也通过向量机回归、反向传播(BP)神经网络、径向基函数(RBF)神经网络对土壤有机质含量和产量进行回归模拟，结果表明土壤有机质和作物产量呈正相关关系。王茵茵<sup>[27]</sup>利用随机森林（RF）算法，基于 AWIFS、MODIS 和 SRTM 遥感数据及实测样点数据，预测了陕西省及其不同地貌区的土壤有机质空间分布状况。胡贵贵<sup>[28]</sup>则对陕西省旬邑县 2008 年和 2018 年苹果区的土壤采样数据和环境变量进行了分析，采用 PCA 和 PLS 方法进行特征选择，并使用随机森林（RF）、支持向量回归（SVR）和 K 近邻算法构建了土壤

养分预测模型。陈道坤<sup>[29]</sup>将土壤 pH、全氮、全磷、全钾、全硫、硫化物、铵态氮、硝态氮、腐殖质共 9 个参数作为模型的输入参数，有机质作为模型的输出参数，提出随机森林回归（RFR）和 BP 神经网络结果加权融合模型（BP-RFR）。

国外也有不少学者对有机质含量进行了预测，例如 Siewert<sup>[30]</sup>使用多元线性回归、人工神经网络、支持向量机和随机森林这四种机器学习算法对瑞典北部的有机质储量进行了预测，并发现随机森林模型表现最佳。Mundada<sup>[31]</sup>使用随机森林、支持向量机、自适应提升和 K 近邻方法对 30m 分辨率下的土壤有机质进行了预测，并使用 R^2 和 RMSE 评估模型，结果显示随机森林的准确性最高。

针对山西省土壤有机质含量的预测中，郑宇桐<sup>[32]</sup>使用了 MLR、GAM、GWR、SGWR 对山西省全省范围的土壤有机质含量进行回归预测。其对环境变量进行了处理筛选，保留了六个环境变量并输入模型，由此对山西省有机质含量进行了预测。高鹏利<sup>[33]</sup>利用地形、植被、土壤属性数据，使用 Boruta 算法筛选特征，并运用多种方法（包括 OK、BPNN、GA-BPNN、GA-BPNN-OK）建立了 SOM 预测模型。

总体而言，随着机器学习算法的成熟，对土壤有机质含量预测的研究逐渐兴起。通过建立机器学习模型，研究者能够深入理解土壤有机质含量背后的影响因素，并提供了一种可行的方法来预测未知情况。然而，目前在预测方向的研究仍然占据较少的比例，特别是在山西省有机质含量的预测方面，相关使用机器学习算法的文献近两年才开始出现。因此，预测方向的研究仍然存在相当的局限性。

## 1.4 研究内容

为解决以上问题，本研究通过三步拆解研究内容：描述现象、理解时空变化——耦合多种因素、定量理解影响——设定未来情景、预测未来。

### （1）山西省农田土壤有机质时空演变特征

基于全国第二次土壤普查数据（1982 年）、山西省耕地质量调查数据（2012 年）和实测数据（2022 年），将山西省划分为晋北、晋中、晋西南、晋东南，用描述性统计、半方差函数、克里格插值等方法，定量分析山西省过去 40 年间农田表层（0-20cm）土壤有机质时空演变特征。

### （2）山西省农田土壤有机质的影响因素

基于全国第二次土壤普查数据（1982 年）、山西省耕地质量调查数据（2012 年）和实测数据（2022 年），结合从 Google Earth Engine 及数据库平台上获取的

气候、地形、农资投入等数据，利用定性分析、分布图可视化及相关性分析的方法，解析土壤有机质与气候（降雨量、温度、潜在蒸发量）、理化指标（pH）、养分指标（全氮、有效磷、速效钾等）、地形（坡度、高程等）、植被（NDVI）、农资投入（化肥）等因素的关系，揭示影响土壤有机质的关键因子。

### （3）山西省农田土壤有机质变化趋势预测

建立多种预测模型，包括线性回归、支持向量机、随机森林、XGBoost 及 LightGBM，通过超参数调优、特征选取的方式，利用多种因子预测有机质含量，并不断优化模型预测结果，比较预测精度。假定 10 年后各因子的变化情形，利用训练好的模型预测山西省农田表层土壤有机质含量。

## 1.5 技术路线

本研究的技术路线如图所示。通过文献调研、实地调研等多种途径收集山西省土壤有机质、氮磷钾、降雨、温度、地形及化肥投入等数据资料，作为本研究的数据基础。在此基础上，首先是摸清山西省有机质分布情况（第三章），通过描述性统计、半方差函数、克里格插值和地图分布可视化，梳理有机质含量以及其他土壤养分的时空分布情况。其次是理解可能的影响因素的时空分布情况（第四章），通过相关性分析、定性分析和可视化，理解土壤养分、地形、气候、植被和农资投入对于土壤有机质含量的影响。然后，通过更为深入的模型，挖掘各类因子对土壤有机质含量的影响（第五章）。构建多元线性回归和多种非线性模型，包括支持向量机、随机森林、XGBoost、LightGBM，并对模型进行调参和特征优化，成功地使用各类影响因子预测有机质含量。最终依据因子趋势和政策，设定未来十年后的情景，基于该情景对于有机质含量进行预测。最后，对本研究进行系统性地总结，分析不足并展望未来。

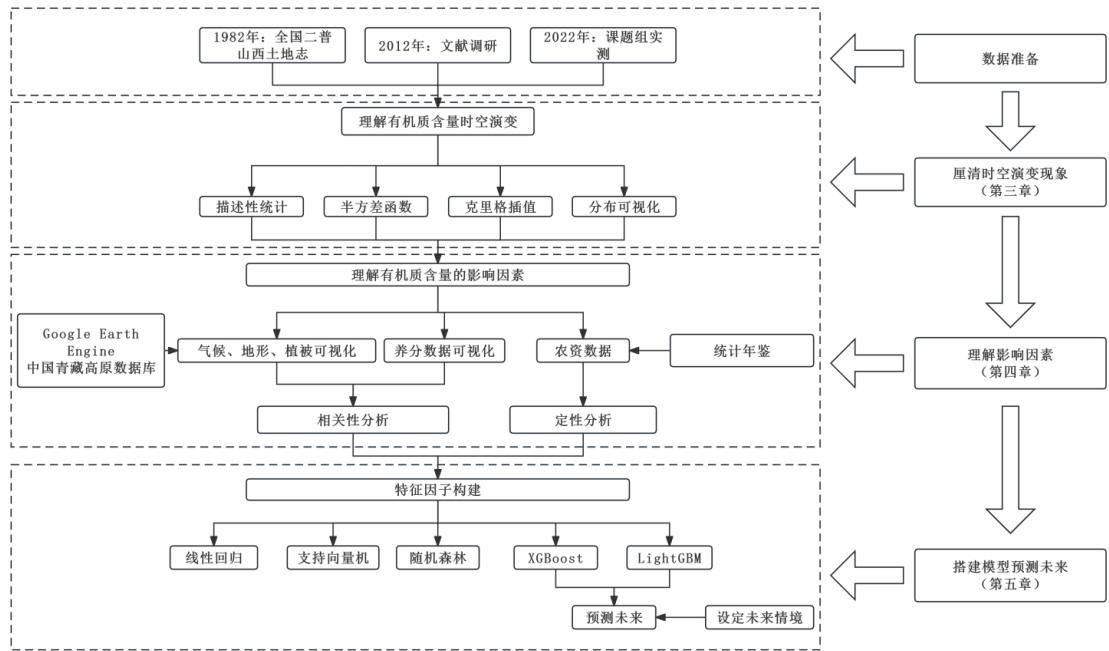


图 1.3 技术路线

## 第2章 数据来源与分析方法

### 2.1 研究区概况

山西省位于华北地区，总面积为 15.67 万平方千米。该省位于中纬度的内陆地带，气候特点是四季分明、雨热同期。东南部属于半湿润温带季风气候，而西北部则属于半干旱温带大陆性气候。山西省的地势主要由黄土覆盖的山地高原组成。高原内部起伏不平，河谷纵横交错，在该省地形中，山地和丘陵占据了约 80% 的比例。



图 2.1 山西省行政区域划分

本研究的焦点是山西省耕地土壤的情况。山西省的耕地总面积为 5804.25 万亩。其中，水田面积为 7.53 万亩，占比 0.13%；水浇地面积为 1571.73 万亩，占比 27.08%；旱地面积为 4225.00 万亩，占比 72.79%<sup>[34]</sup>。

山西省可以划分为晋北、晋中、晋西南和晋东南。晋北包括大同市、朔州市

和忻州市。晋中包括太原市、吕梁市、阳泉市和晋中市。晋西南包括临汾市和运城市。晋东南地区包括长治市和晋城市。

表 2.1 山西省行政区域划分及耕地面积与样点概况

区域	市	1982 年 样点数量 (个)	2012 年样点 数量 (个)	2022 年样点 数量 (个)
晋北	大同市、朔州市、 忻州市	66	176	19
晋中	太原市、吕梁市、 阳泉市、晋中市	50	250	13
晋西南	临汾市、运城市	51	99	15
晋东南	长治市、晋城市	26	120	5

## 2.2 数据来源

### 2.2.1 土壤数据来源

本研究所使用的土壤数据，为三个年份的数据，1982 年、2012 年和 2022 年。其中 1982 年数据来源于全国二普数据中的记载。2012 年数据来源于山西省各县市耕地地力评价与利用，通过提取、整理文献中的数据整理为数据集。2022 年数据来源于课题组前往山西进行实测时获得的实测数据。

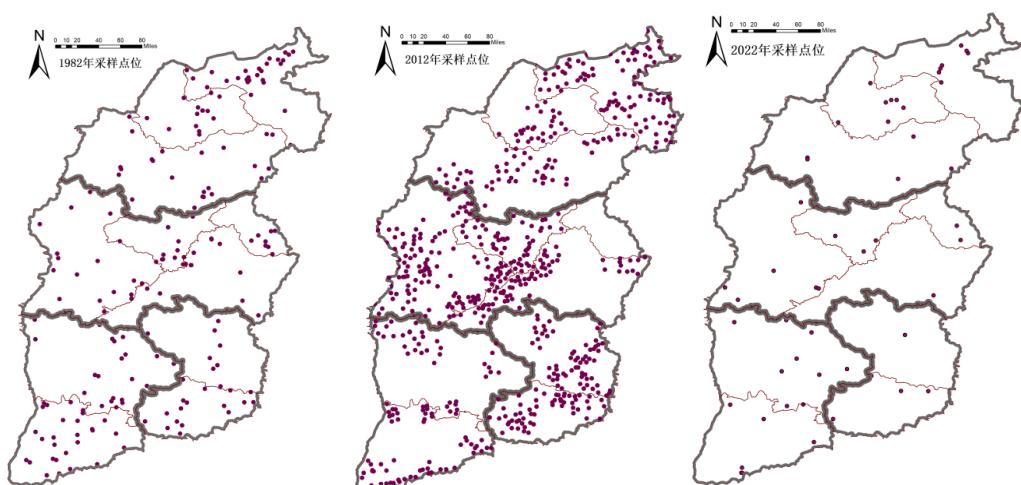


图 2.2 不同年份土样采集点位分布

由于数据来源不同，数据的格式和记录方式也存在较大差异。因此，有必要统一数据标准，以确保后续分析的有效性。具体的数据处理步骤如下：第一，针对同一采样点，将不同土层深度的土壤养分数据取平均值，作为该点位的代表性数据。第二，对于同一乡镇范围内的多个采样点，也采用取平均值的方式，以代表该乡镇的整体情况。第三，由于数据记录方法的差异，需要将采样点的地址信息细化到市、区、乡镇三个层级，以便于后续的空间分析。此外，由于 1982 年至 2022 年期间，山西省的行政区划发生了一些变化，因此还需要将历史数据中的地名更新为当前的行政区划信息，确保数据的时间连续性。由于 1982 年和 2012 年的原始数据缺乏经纬度信息，仅有类似“三楼乡上宅村”这样的地点描述，需要通过人工查询的方式，以距离该村行政所在地中心位置最近的农田地理坐标作为样点坐标。同时，还需要关注部分村庄名称的变更情况。数据整理完成后，按照同一经纬度坐标作为一个采样点，统计了各年份的数据覆盖情况。2022 年样点地理坐标信息来源于奥维地图，为实采样点位置。结果显示，1982 年共有 209 个采样点，2012 年为 641 个采样点，2022 年为 53 个采样点。

### 2.2.2 气候、地形、植被数据来源

为研究非土壤养分因子对有机质含量的影响情况，对气候、地形、植被数据进行搜集。

第一，气候因子共 3 个，包含年总降水量、年均温、年潜在蒸发量。总降水数据来源于《中国 1km 分辨率逐月降水量数据集（1901-2022）》<sup>[35]</sup>，对于 1982、2012、2022 年的逐月降水量分别按年加和，得到三年分别的年总降水量。年均温数据也来源于《中国 1km 分辨率逐月平均气温数据集（1901-2022）》<sup>[36]</sup>，对于这三年的数据做年平均，得到年均气温。年潜在蒸发量（PET）指的是土壤蒸发的耗水量以及植物蒸腾作用所导致的的总耗水量之和，通常使用一些估算方法得到。数据来源于国家青藏高原数据中心的《中国 1km 分辨率逐月潜在蒸散发数据集（1901-2022）》<sup>[37]</sup>。此数据集中的 PET 是由 Hargreaves 计算式得到的，公式如下：

$$PET = 0.0023 \times S0 \times \sqrt{MaxT - MinT} \times (MeanT + 17.8) \quad (2-1)$$

其中 MaxT、MinT、MeanT 分别为最高、最低、和平均温度。S0 是到达大气层顶的理论太阳辐射。

以上数据导出为 nc 格式，在 ArcMap 中转换为栅格 tiff 格式后，便可在地图

上进行可视化，绘制山西省整体图像。此外，还需要提取之前土壤采样点的数据，用于后续搭建模型，寻找有机质和各影响因子之间的关系。使用 ArcMap 的多值提取到点功能，对采样点位的气候数据进行提取。

第二，地形因子共 3 个，包含高程(Elevation)、坡度(Slope)、坡向(Aspect)。数据均来源于微软公司的 Google Earth Engine 平台。GEE 平台集成了例如 Landsat、MODIS、Sentinel 等卫星的巨量历史数据，有大量的遥感影像供用户使用。GEE 强大的计算力使得用户可以在其平台上通过编写代码获得数据。本研究使用与该平台交互的代码编辑器（基于网页的集成开发环境 IDE），编写和运行使用 Earth Engine JavaScript API 的脚本。由此获得所需要的山西省数据，将 tiff 数据下载后按照气候数据的类似步骤对原始数据进行可视化和提取。

第三，植被数据共 1 个，即是最大归一化植被指数(NDVI)。NDVI 随着植被覆盖增多而增加。而植被的生长情况与土壤营养物质的含量是有非常强烈的相关性的。因此 NDVI 在土壤有机质预测中常常起到重要作用。NDVI 的计算公式由近红外波段 NIR 和红波段 R 计算得到<sup>[38]</sup>，公式为：

$$NDVI = (NIR - R) / (NIR + R) \quad (2-2)$$

本研究 NDVI 的数据来源为 ORNL DACC<sup>[39]</sup>。同理下载 nc 数据后，在 ArcMap 中完成可视化，并且提取采样点位的数据。

### 2.2.3 农资数据来源

农资数据体现为灌溉、施肥数据。数据来源于中国统计年鉴记载。时间范围涵盖 1998 年至 2022 年，数据包括耕地灌溉面积和农业化肥施用量。农业化肥施用量中包含氮肥、磷肥、钾肥、复合肥分别的施用量，以及总化肥施用量。农资数据的空间颗粒度较粗糙，仅有山西省作为整体的数据，因此这部分农资数据仅用于在第四章中分析对有机质含量的影响，不参与第五章的机器学习模型搭建。

## 2.3 研究方法

### 2.3.1 地统计学理论

经典统计学中常用的参数(如最小值、最大值、平均值、标准差等)虽然能够反映数据的基本分布情况，但却无法联系地理信息，无法进行深入的地理数据分析。为此，法国统计学家 Matheron 教授在 20 世纪 60 年代提出了地统计学的概念<sup>[40]</sup>。地统计学的主要内容包括协方差函数、变异函数、区域化变量和克里金插值。

区域化变量理论是地统计学的基础，用于描述变量的空间分布。实际应用中，一些例如土壤数据、温度、降雨等和空间分布有关的现象都可以用区域化变量进行表达。半方差函数是专用于描述区域化变量的函数，也是克里金插值的基础，计算公式如下：

$$r(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \quad (2-3)$$

半方差函数图，是 $r(h)$ 作为距离 $h$ 的函数的图形，有四个关键指标。变程反映了区域化变量影响范围的大小。块金值反映了区域化变量内部随机性的程度。基台值则表示了变量变化的幅度。基台比则反映了变量的空间变异程度，较高的数值表示随机部分引起的空间异质性程度较高，较低的数值则表示由空间自相关部分引起的空间变异较大。当基台比小于25%时，说明变量具有强烈的空间相关性；而在25%至75%之间，代表中等程度的空间相关性；而大于75%则代表变量的空间相关性很弱。

对于半方差函数拟合函数进而建立变异函数理论模型。地统计学中常用的模型有球状模型、指数模型、高斯模型、线状模型等。在本研究中，使用平均标准误差来衡量不同模型的准确程度。最终选择平均标准误差最小的模型作为最终的理论模型。例如，球状模型公式为：

$$r(h) = \begin{cases} 0 & h = 0 \\ C_0 + C \left[ \frac{3h}{2a} - \frac{1}{2} \left( \frac{h}{a} \right)^3 \right] & 0 < h \leq a \\ C_0 + C & h > a \end{cases} \quad (2-4)$$

高斯模型的公式为：

$$r(h) = \begin{cases} 0 & h = 0 \\ C_0 + C[-\exp(-h/a)^2] & 0 < h \leq \sqrt{3}a \\ C_0 + C & h > \sqrt{3}a \end{cases} \quad (2-5)$$

指数模型的公式为：

$$r(h) = \begin{cases} 0 & h = 0 \\ C_0 + C[1 - \exp(-h/a)] & 0 < h \leq 3a \\ C_0 + C & h > 3a \end{cases} \quad (2-6)$$

以上地统计学分析为克里格插值做好了准备。克里金插值法能结合空间属性，给出理论统计的插值误差，还能结合采样数据反映的区域化变量以及样本点的空间位置，对区域化变量的取值进行线性无偏最优估计，其公式为：

$$\hat{z}_o = \sum_{i=1}^n \lambda_i z_i \quad (2-7)$$

其中  $\hat{z}_o$  是点  $(x_0, y_0)$  处的估计值。 $\lambda_i$  是权重系数。权重系数满足方差最小和无偏估计，即：

$$\min_{\lambda_i} \text{Var}(\hat{z}_o - z_0) \quad (2-8)$$

$$E(\hat{z}_o - z_0) = 0 \quad (2-9)$$

$\hat{z}_o$  是  $z_0$  的最优线性无偏估计量。在克里格插值之前，需保证数据符合正态分布，否则要对数据进行变换。常见变换如对数 (log) 变换、Box-Cox 变换。

对数变换即是对数据取对数，对数变换后符合正态分布的数据符合对数正态分布。Box-Cox 变换的一般形式如下公式所示，其中  $\lambda$  为变换参数，可以用极大似然估计解得  $\lambda$  值。

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases} \quad (2-10)$$

### 2.3.2 相关性分析

在本研究中，使用皮尔逊相关系数来衡量两个连续随机变量之间的相关性，计算公式如下：

$$\begin{aligned} \rho_{X,Y} &= \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \\ &= \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} \\ &= \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \sqrt{E(Y^2) - E^2(Y)}} \end{aligned} \quad (2-11)$$

其中  $E$  是数学期望， $\text{cov}(X, Y)$  是协方差， $\sigma$  是标准差。使用  $t$  检验来验证 Pearson 相关系数的显著性。

原假设:  $H_0: \rho = 0$ ,  $= \frac{r-0}{S_r} = \frac{\sqrt{(n-2)r}}{\sqrt{(1-r^2)}}$ , 自由度为  $n - 2$ ,  $r$  为样本相关系数。表达显著性的 p 值公式如下:

$$p(|t| \geq t_{1-\frac{\alpha}{2}}(n-2)) \quad (2-12)$$

将 p 值与置信水平  $\alpha$  进行比较 (一般取 0.05)。当 p 小于  $\alpha$  时, 即认为否定原假设, 结果显著。在结果显著的前提下, 可以通过相关系数的绝对值范围来判断变量的相关强度。其中 0.8-1.0 表达变量有极强相关, 0.0-0.2 表达变量间为极弱相关甚至无相关。

### 2.3.3 多元线性回归

多元线性回归即是将 Y 变量与多个 X 变量建立线性回归关系。其满足以下关系: 假设有 n 个观测值,  $i = 1, 2, \dots, n$ , 满足:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_m X_{im} + \varepsilon_i \quad (2-13)$$

其中,

$$\varepsilon_i | X_i \sim i.i.d. N(0, \sigma^2) \quad (2-14)$$

当  $m=1$  的时候, 即是简单线性回归模型。其中  $\beta_0$  为截距,  $\beta_j$  为回归系数,  $\varepsilon_i$  为误差项。回归系数  $\beta_j$  的意义代表着在剔除其他 X 变量的影响下,  $X_j$  单位变化对 Y 的影响。通过多元线性回归可以得到不同 X 对于 Y 的影响, 可以通过观察系数和显著性 p 值理解影响程度。

### 2.3.4 支持向量机

支持向量机回归把数据点映射到高维空间中, SVR 的目标是令所有的数据点在某个精度  $\varepsilon$  内拟合  $f(x) = \mathbf{w}^T \mathbf{x}_i + b_i$ , 同时令方程最平缓, 满足  $\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2$ 。因此总结 SVR 为以下数学表达:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad (2-15)$$

$$\text{s.t. } \begin{cases} y_i - (\mathbf{w}^T \mathbf{x}_i + b_i) \leq \varepsilon, \\ (\mathbf{w}^T \mathbf{x}_i + b_i) - y_i \leq \varepsilon, \end{cases} \quad 1 \leq i \leq n. \quad (2-16)$$

### 2.3.5 随机森林

随机森林是基于多个决策树的集成学习算法。该算法创建 bootstrap 数据集,

有放回地构建多个子数据集，在构造决策树的时候随机选取部分特征，对这些特征计算分裂后的收益，决定每棵树的每个分裂特征节点。最终多个树的回归结果取平均，作为最终的预测结果。

### 2.3.6 XGBoost

XGBoost 是基于梯度提升决策树，进一步优化后的模型。它也是多个决策树的集合。XGBoost 算法的目标函数公式为<sup>[41]</sup>:

$$\begin{aligned}\tilde{l}^{(t)} &\simeq \sum_{i=1}^n \left[ g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right] + \Omega(f_i) \\ &= \sum_{i=1}^n \left[ g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \\ &= \sum_{j=1}^T \left[ \left( \sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left( \sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T\end{aligned}\quad (2-17)$$

其中， $I_j = \{i \mid q(x_i) = j\}$  表示叶子  $t$  的实例集。

$$g_i = \frac{\partial l(\hat{y}_i^{(t-1)}, y_i)}{\partial \hat{y}_i^{(t-1)}} \quad (2-18)$$

$$h_i = \frac{\partial^2 l(\hat{y}_i^{(t-1)}, y_i)}{\partial (\hat{y}_i^{(t-1)})^2} \quad (2-19)$$

可以计算出叶子  $j$  的最优权重  $w_j^*$ ，目标函数的最优解为:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (2-20)$$

$$l^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^T \frac{\left( \sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \quad (2-21)$$

### 2.3.7 LightGBM

LightGBM 是基于决策树算法的分布式梯度提升框架，使用直方图算法将特征离散化，不像往常的 GBDT 模型的按层生长的决策树生长策略，LightGBM 使用带深度限制叶子生长策略，对分裂增益最大的叶子分裂，更快且精度更高。

## 第3章 山西省农田土壤有机质时空演变特征

### 3.1 山西省农田土壤有机质时空变化特征

山西省整体有机质含量平均值 1982 年为 9.4g/kg, 2012 年为 15.14g/kg, 2022 年为 17.59g/kg。1982-2012 年增速为 61%, 2012-2022 年增速为 16.2%。

1982 年到 2012 年, 山西省 11 各市的有机质含量呈现出了不同程度的上升趋势。平均 2012 年的有机质含量相较 1982 年增加了 59%, 增长幅度很大。增速最高的是运城市, 达到了 149% 的增速。增速最低的是忻州市, 为 2.4%。从 2012 年到 2022 年, 增速放缓, 有三个市出现了负增长, 大同市为-11.6%, 朔州市为-5.4%, 忻州市为-17.0%。整体上平均增速为 29.6%, 低于 1982-2012 年的 59%, 增速仅 1982-2012 的一半。但在其中也有表现很好的市, 例如晋中市增速达到了 125%。

表 3.1 1982-2022 年山西省各市农田土壤有机质含量 (单位: g/kg)

年份	大同市	晋城市	晋中市	临汾市	吕梁市	朔州市	太原市	忻州市	阳泉市	运城市	长治市
1982	8.00	12.72	9.32	10.47	6.49	8.76	15.39	10.92	12.91	6.73	11.16
2012	13.80	20.92	16.82	13.51	12.00	10.87	18.45	11.19	ND	16.73	18.78
2022	12.20	28.07	37.82	22.85	13.75	10.28	27.98	9.28	26.47	17.45	24.75

(ND: 数据缺失)

### 3.2 山西省农田土壤养分时空变化特征

全氮含量平均值 1982 年为 0.627g/kg, 2012 年为 0.853g/kg, 2022 年为 0.931g/kg。有明显增长, 1982-2012 年增速为 36%, 2012-2022 年增速为 9%。pH 值仅有 1982 年和 2022 年的数据, 1982 年 pH 平均值为 8.07, 2022 年为 8.26, 增长率为 2.46%。全磷数据只有 1982 年和 2022 年的数据, 1982 年山西省全磷含量平均值为 0.56g/kg, 2022 年山西省全磷含量平均值为 0.82g/kg。1982-2022 年相对增长了 45.20%。有效磷数据仅有 2012 与 2022 年的数据。2012 年山西省有效磷含量平均值为 11.96mg/kg, 2022 年山西省有效磷含量平均值为 21.01mg/kg。1982-2022 年相对增长了 75.65%。速效钾数据仅有 2012 与 2022 年的数据。2012 年山西省速效钾

含量平均值为 142.30mg/kg, 2022 年山西省速效钾含量平均值为 163.76mg/kg。2012-2022 年相对增长了 15.08%。缓效钾数据仅有 2012 年的数据。2012 年山西省缓效钾含量平均值为 784mg/kg。碱解氮数据仅有 2022 年的数据。2022 年山西省缓效钾含量平均值为 67.48mg/kg。

表 3.2 土壤肥力指标时间演变

区域	市级	年份	pH	全氮 g/kg	全磷 g/kg	有效磷 mg/kg	速效钾 mg/kg	缓效钾 mg/kg	碱解氮 mg/kg
晋北	大同市	1982	8.4	0.5	0.56	ND	ND	ND	ND
		2012	ND	0.78	ND	9.44	115.48	743.48	ND
		2022	8.06	0.61	0.47	5.96	113.19	ND	43.42
	朔州市	1982	8.4	0.54	0.5	ND	ND	ND	ND
		2012	ND	0.56	ND	9.45	94.36	590.8	ND
		2022	7.93	0.54	0.34	10.2	101.9	ND	43.27
		1982	8.4	0.54	0.5	ND	ND	ND	ND
	朔州市	2012	ND	0.56	ND	9.45	94.36	590.8	ND
		2022	7.93	0.54	0.34	10.2	101.9	ND	43.27
	均值	1982	8.4	0.53	0.52	ND	ND	ND	ND
		2012	ND	0.63	ND	9.44	101.4	641.69	ND
		2022	7.97	0.56	0.38	8.78	105.66	ND	43.32
晋中	太原市	1982	8.01	0.66	0.55	ND	ND	ND	ND
		2012	ND	0.89	ND	11.8	134.44	774.07	ND
		2022	8.2	1.3	1.16	23.45	192.3	ND	91.81
	阳泉市	1982	7.77	0.74	0.76	ND	ND	ND	ND
		2012	ND	ND	ND	ND	ND	ND	ND
		2022	8.6	1.02	0.8	26.34	179.1	ND	56.96
	晋中市	1982	8.02	0.58	0.54	ND	ND	ND	ND
		2012	ND	0.97	ND	14.48	165.15	669.77	ND
		2022	7.5	1.29	0.67	15.86	105.85	ND	95.79

续表 3.2 土壤肥力指标时间演变

区域	市级	年份	pH	全氮 g/kg	全磷 g/kg	有效磷 mg/kg	速效钾 mg/kg	缓效钾 mg/kg	碱解氮 mg/kg
晋中	吕梁市	1982	8.03	0.45	0.63	ND	ND	ND	ND
	吕梁市	2012	ND	0.63	ND	8.45	129.78	745.17	ND
	吕梁市	2022	8.88	0.77	0.84	24.7	107.87	ND	52.18
	均值	1982	7.9575	0.6075	0.62	ND	ND	ND	ND
	均值	2012	ND	0.83	ND	11.57	143.12	729.67	ND
	均值	2022	8.295	1.10	0.86	22.58	146.28	ND	74.185
晋西南	临汾市	1982	8.21	1.06	0.5	ND	ND	ND	ND
	临汾市	2012	ND	0.76	ND	11.43	163.46	867.69	ND
	临汾市	2022	8.48	1.31	1.26	20.06	147.43	ND	93.87
	运城市	1982	8.23	1.4	0.56	ND	ND	ND	ND
	运城市	2012	ND	0.86	ND	15.97	179.72	962.54	ND
	运城市	2022	8.76	1.03	1.03	38.41	266.91	ND	74.14
晋东南	均值	1982	8.22	1.23	0.53	ND	ND	ND	ND
	均值	2012	ND	0.81	ND	13.7	171.59	915.12	ND
	均值	2022	8.62	1.17	1.15	29.24	207.17	ND	84.01
	晋城市	1982	7.79	0.81	0.42	ND	ND	ND	ND
	晋城市	2012	ND	1.36	ND	14.61	166.96	794.6	ND
	晋城市	2022	8.22	1.47	0.9	26.39	285.57	ND	74.64
晋东南	长治市	1982	7.9	0.61	0.45	ND	ND	ND	ND
	长治市	2012	ND	0.98	ND	9.71	169.01	976.47	ND
	长治市	2022	8.53	1.27	0.71	25.82	206.86	ND	72.78
	均值	1982	7.845	0.71	0.435	ND	ND	ND	ND
	均值	2012	ND	1.17	ND	12.16	167.99	885.54	ND
	均值	2022	8.38	1.37	0.81	26.11	246.22	ND	73.71

(ND: 数据缺失)

### 3.3 山西省耕地土壤肥力指标的时空分布特征

#### 3.3.1 耕地有机质的时空空间分布情况

使用 Arcgis 建模，对于 1982 年、2012 年、2022 年三年的有机质数据分别建立克里格插值模型。首先对于 1982 年的有机质数据进行建模。

表 3.3 1982 年有机质描述性统计

点位数	平均值	最小值	最大值	标准差	偏度	峰度	中位数	上分位数	下分位数
193	9.40	2.16	49.86	7.07	2.83	12.80	6.90	5.40	10.64

通过绘制直方图可以发现，数据并未呈现出正态分布的特征。这意味着直接将其纳入克里格插值模型可能会产生偏差。为此，采取相应的数据变换措施以满足插值模型的基本假设。尝试了对数变换和 Box-Cox 变换两种方法，对比分析了变换后数据的分布情况，最终选择对数变换作为数据预处理手段。

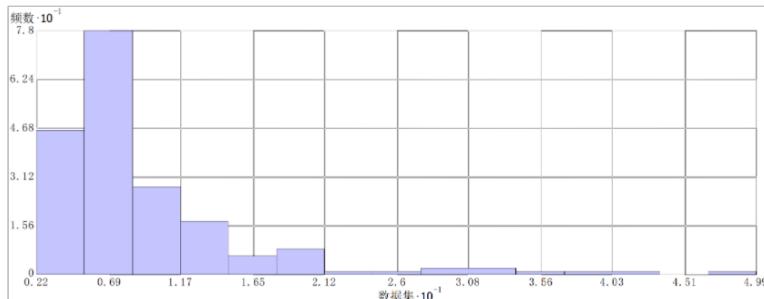


图 3.1 1982 有机质数据分布

对于 1982 年耕地有机质数据变换前后的正态 QQ 图如下图所示，可以看出经过对数变换后的有机质数据更加形似正态分布。

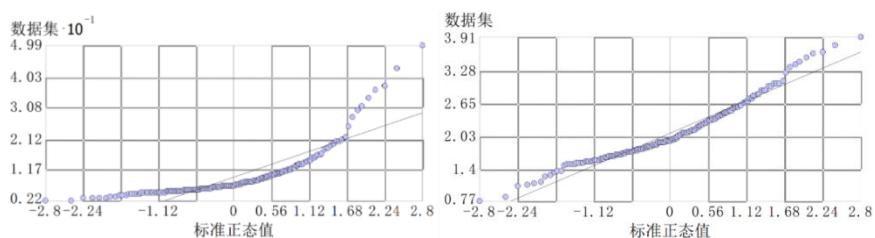


图 3.2 1982 年变换前（左）后（右）有机质数据正态 QQ 图

在变换的基础上，建立克里格模型。克里格插值建立在半方差函数的基础之上，半方差函数可以被某些理论模型所拟合，如三角、球面、指数、高斯和稳定模型等。1982 年不同模型的拟合平均标准误差见下表，稳定模型是平均标准误差最小的模型。

表 3.4 1982 有机质不同半方差函数拟合结果

理论模型	平均标准误差
三角	4.7771
球面	4.7722
指数	4.773
高斯	4.7885
稳定	4.77162

稳定性理论模型拟合得到的半方差函数表示图如下图所示。块金值为 0.1527498，偏基台值为 0.1764477，基台值为 0.3291975，块基比为 46.4%。此处的块基值为 46.4%，介于 25%~75% 之间，说明 1982 年山西省耕地有机质含量具有中等程度的空间相关性。

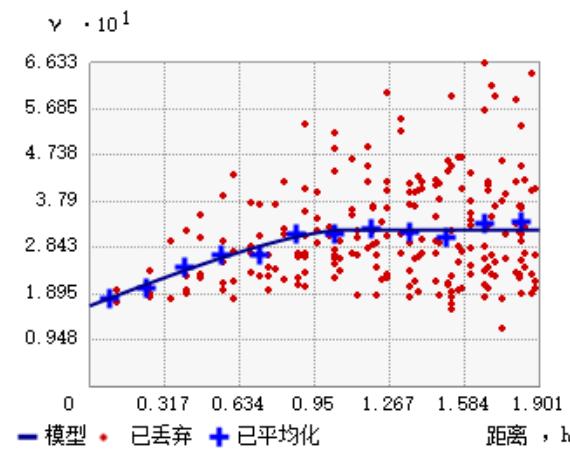


图 3.3 1982 年山西省耕地有机质半方差函数

经过前述的数据预处理和半方差函数模型的拟合，已经建立了 1982 年土壤有机质的空间插值分析框架。接下来即是利用该模型及已有的点位观测数据，通过普通克里格插值法对整个山西省范围内的有机质分布状况进行估算和制图。

接下来对于 2012 年数据进行建模，其基本分布情况如下表所示。

表 3.5 2012 年有机质描述性统计

点位数	平均值	最小值	最大值	标准差	偏度	峰度	中位数	上分位数	下分位数
645	15.14	4.1	50.98	6.32	1.06	4.99	14.05	10.57	18.59

进行对数变换后数据符合正态分布，QQ 图如下所示。

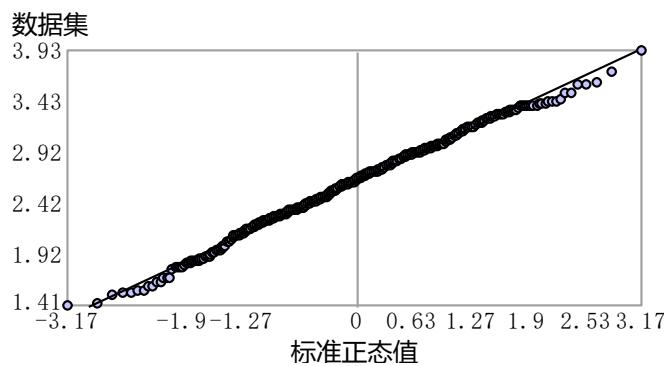


图 3.4 2012 年变换后有机质数据正态 QQ 图

使用不同半方差函数理论模型拟合数据，指数模型的误差最小，选择指数模型作为理论模型。

表 3.6 2012 有机质不同半方差函数拟合结果

理论模型	平均标准误差
三角	4.48426
球面	4.42575
指数	3.99881
高斯	4.63509
稳定	4.0488

指数理论模型拟合得到的半方差函数表示图如下图所示。块金值为 7.674，偏基台值为 34.2328，基台值为 41.9068，块基比为 18.31%，小于 25%，2012 年山西省耕地有机质含量具有很强烈的空间相关性。

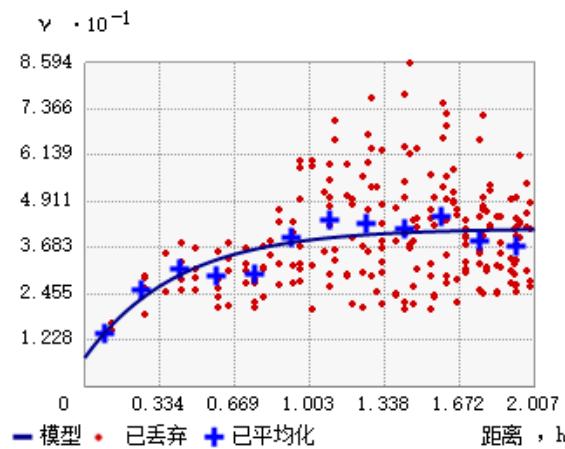


图 3.5 2012 年山西省耕地有机质半方差函数

就 2022 年数据而言，描述性统计分析结果如下所示。由于该年份的观测点位数量较为有限，仅为 52 个，这可能会影响后续建模的精度和可靠性。

表 3.7 2022 年有机质描述性统计

点位数	平均值	最小值	最大值	标准差	偏度	峰度	中位数	上分位数	下分位数
52	17.59	5.41	47.29	9.80	0.93	3.18	14.52	10.17	25.04

同理，拟合不同模型，发现高斯型最为合适。块金值 30.75，偏基台值 95.865，基台值为 126.615，块基比为 24%，2022 年山西省耕地有机质含量具有很强烈的空间相关性。

表 3.7 2022 有机质不同半方差函数拟合结果

理论模型	平均标准误差
三角	6.58464
球面	6.6116
指数	6.9855
高斯	6.5367
稳定	6.540

下图为 2022 年耕地有机质半方差函数图。

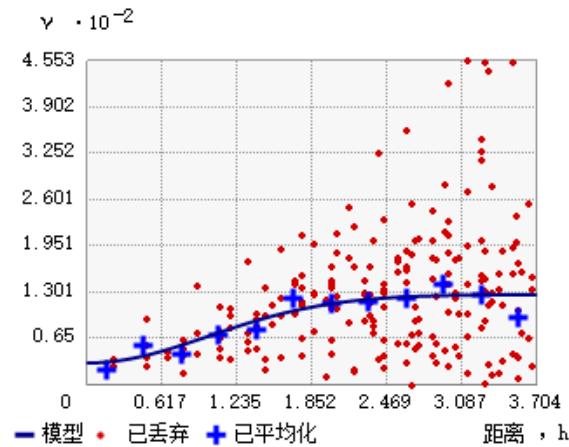


图 3.6 2022 年山西省耕地有机质半方差函数

通过完成 1982 年、2012 年和 2022 年三个时期的土壤有机质半方差函数建模，可以清晰地观察到 2012 年和 2022 年两个时期的块金效应具有非常强烈的空间相关性。这表明，在这些年份的有机质空间变异中，起主导作用的是结构性因素，如气候、土壤母质、土壤类型和地形等。相比之下，诸如种植制度、施肥等随机因素的影响相对较小。接下来将模型利用于克里格插值中，分别对三年的耕地土壤有机质含量进行插值。

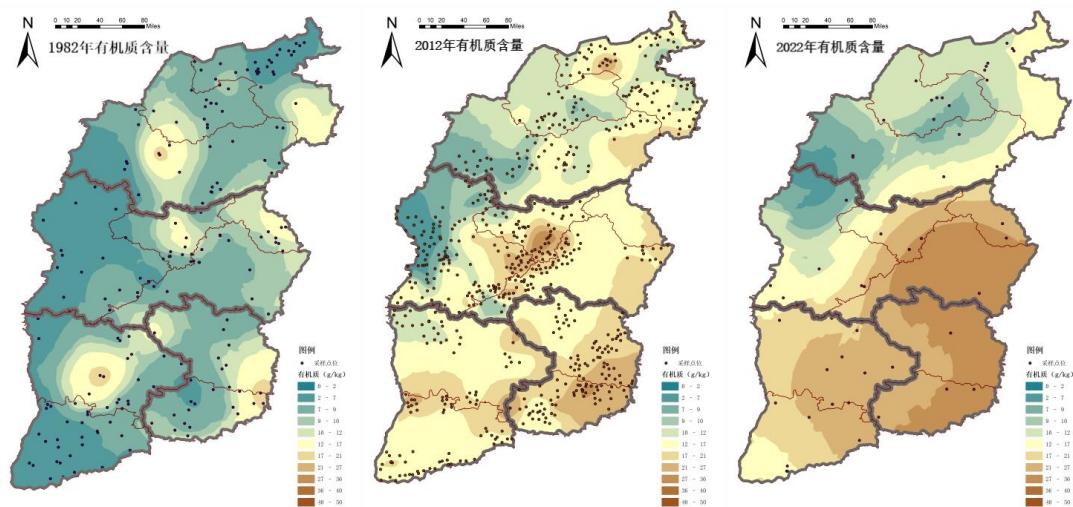


图 3.7 山西省不同年份农田土壤有机质含量分布

经过对 1982 年、2012 年和 2022 年三个时期山西省土壤有机质空间分布的分析，我们可以清楚地观察到其时空演变特征：1982 年，全省平均有机质含量较低，仅 9.4064 g/kg。整体呈现由西向东逐渐增加的趋势，晋东南地区含量相对较高，而晋西南最南端含量最低。到 2012 年，空间分布格局发生一定变化。晋北和晋中最西侧有机质含量最低，但与 1982 年相比，最南端和最北端已有明显提升。整体呈现由西偏北向东南增加的趋势，晋中心和晋东南地区含量最高。到 2022 年，分布趋势与 2012 年相似，仍以西北向东南递增为主。不过，东南地区有机质含量较 2012 年有所提高，高含量区域呈现扩大趋势。晋东南地区含量最高，晋北和晋中西侧最低。为进一步探讨不同区域的时空变化特点，我们将绘制晋北、晋中、晋西南和晋东南四个亚区域 1982 年、2012 年和 2022 年有机质含量变化图，以更精细地认识各地区的差异性，为针对性的土壤管理和农业生产实践提供参考。

对于晋北地区，1982 年晋北地区整体呈现较低的有机质含量，颜色较深绿，大多为 7g/kg 以下。在晋北地区内部，有机质含量相对均匀。2012 年晋北地区的有机质含量有所增加，但仍然较低，呈现较浅绿色和浅黄色。晋北的西南地区的有机质含量最低，东部地区含量稍高。2022 年晋北地区有机质含量继续增加，但仍然相对较低，呈现淡绿色和浅黄色。西侧地区偏低，东部沿线有机质含量相对较高。

对于晋中地区，1982 年晋中整体有机质含量较低，西侧呈现大片深绿色，有机质含量小于 2g/kg，东侧地区相对含量较高，但总体仍然含量很低。呈现由西向东递增的大致趋势。到了 2012 年，有机质含量有明显提升，大片地区呈现黄色。晋中的中部地区有机质含量最高，西侧地区含量仍然很低。整体呈现从中部向两侧递减的趋势，但是东侧边缘有机质含量偏离趋势，接近中部含量，含量较高。在 2022 年，晋中有机质含量整体发生了明显提升，同时也呈现出了明显的空间分布特征。从晋中地区的西北部到东南部，有机质含量逐渐递增。西北部的地区有机质含量较低，呈现较深绿色；而东南部的地区有机质含量相对较高，呈现深棕色。沿着这个梯度，有机质含量从低到高逐渐增加。

对于晋西南地区，1982 年整体含量很低，大量区域有机质含量仅在 2g/kg 以下。呈现由中部向周围递减的趋势，南端有机质含量最低。2012 年整体有机质含量有大幅提升，中部有一条从西向东贯通的区域有机质含量最高，向两侧递减。2022 年有机质含量再次大幅提升，呈现棕黄色，呈现由东向西递减的趋势。

晋东南地区在 1982 年有机质含量较低，但是含量水平高于同期的其他三个地区，呈现东部高西部低的态势。到了 2012 年，整体有机质含量出现不少提升，

仍然呈现东高西低的分布趋势。2022 年有机质含量再次出现大幅提升，高于 2022 年其他三个地区的水平，大片区域高于 40g/kg 含量。空间分布仍是东多西少。

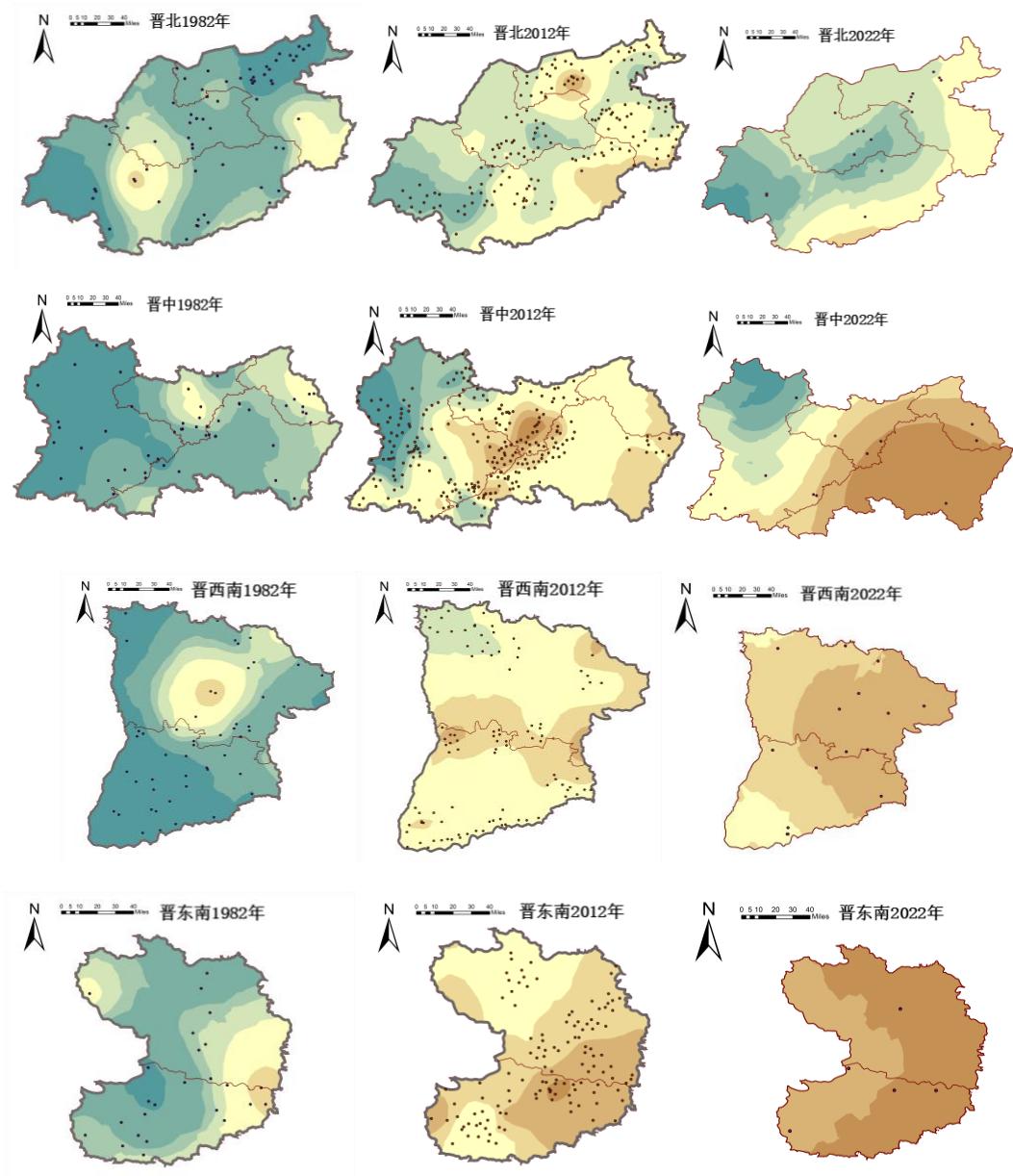


图 3.8 山西省不同区域不同年份有机质含量变化图

### 3.3.2 耕地全氮的时空空间分布情况

从描述性统计数据可以发现，从 1982 年到 2022 年，全氮含量整体有所上升，这体现在 1982 年至 2022 年的全氮含量平均值和中位数发生了提升。其中最为明显的是 1982 年至 2012 年，平均值和中位数发生了大幅度的提升，而 2012 年至 2022 年提升的速度有所放缓。

表 3.8 山西省耕地全氮含量描述性统计

年份	点位数	平均值	最小值	最大值	标准差	偏度	峰度	中位数	上分位数	下分位数
1982	193	0.62	0.15	10.82	0.80	10.91	138.38	0.49	0.39	0.66
2012	647	0.85	0.28	4.58	0.31	3.09	34.22	0.8	0.65	1.02
2022	52	0.93	0.33	1.907	0.44	0.70	2.49	0.81	0.567	1.18

观察数据的分布，发现原始数据基本符合对数正态分布，因此在后续建模进行克里格插值时对原始数据进行对数变换。接着对每年的数据拟合半方差函数，分别根据误差选取最合适的理论模型进行拟合。三年选取的理论模型和平均标准误差、块金值、基台值、偏基台值、块基比、空间相关强度总结于下表。

表 3.9 山西省耕地三年全氮克里格插值结果

年份	模型	平均标准误	块金值	基台值	偏基台值	块基比	空间相关性
1982	三角函数	0.35	0.06	0.25	0.19	24.97%	强
2012	指数函数	0.16	0.02	0.13	0.11	16.29%	强
2022	高斯函数	0.26	0.05	0.36	0.31	15.53%	强

三年山西省整体全氮含量分布随时间变化如下图所示。

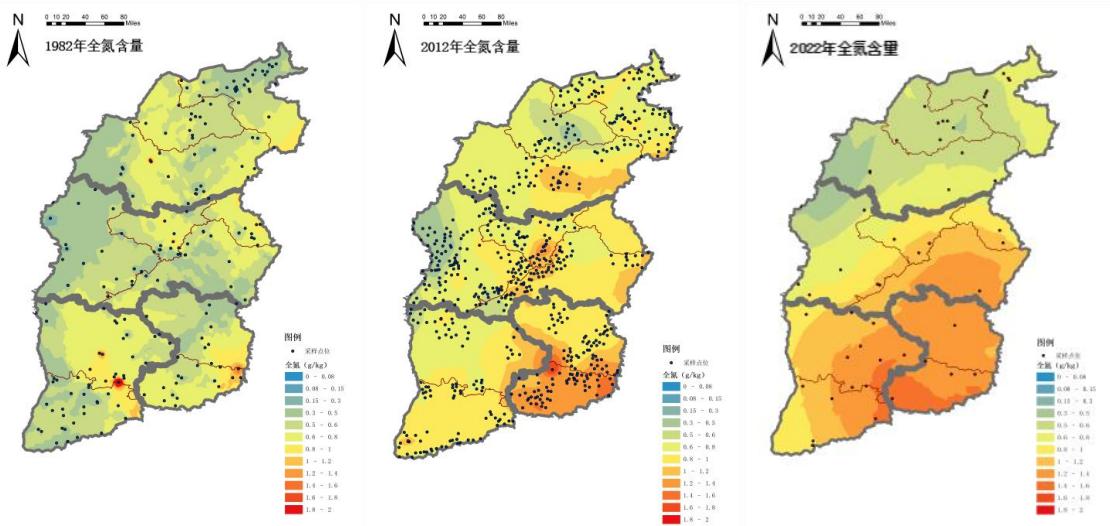


图 3.9 山西省农田土壤不同年份全氮含量变化

1982 年西侧全氮含量最低，中部和北部稍高，但整体平均氮含量仍然偏少，分布相对均匀。2012 年数据点很多，因此插值出来的结果也更加精确。西部有局部的含量仍然较低，相比 1982 年没有大量提升，但是其他地区的全氮含量都有所提升，尤其是东南地区，全氮含量达到 1.4 以上，整体上呈现的趋势是西低东南高，其他地区分布较为均匀。2022 年的采样点位较少，因此插值出的结果并不是很准确，但是整体趋势自西北方向向东南方向递增。西北方向的全氮含量相比于 2012 年有所降低，不排除采样点位过少导致的偏差。

接下来，拆分晋北、晋中、晋西南、晋东南进行细节观察。绘制以下晋北、晋中、晋西南、晋东南（从左至右）1982、2012、2022（从上到下）全氮含量变化图。

出晋北和晋中地区的含量一直相对于晋西南和晋东南偏少。晋北地区自 1982 到 2022 年，全氮含量经历了先增大又减小的趋势，但可能是由于 2022 年点位过少造成的偏差。晋中地区的全氮含量不断提升，每年的趋势都是西侧向东侧有递增趋势。晋西南地区的全氮含量也明显递增，整体空间分布北低南高。晋东南地区的全氮含量随年增长的，空间分布趋势由西向东递增趋势。

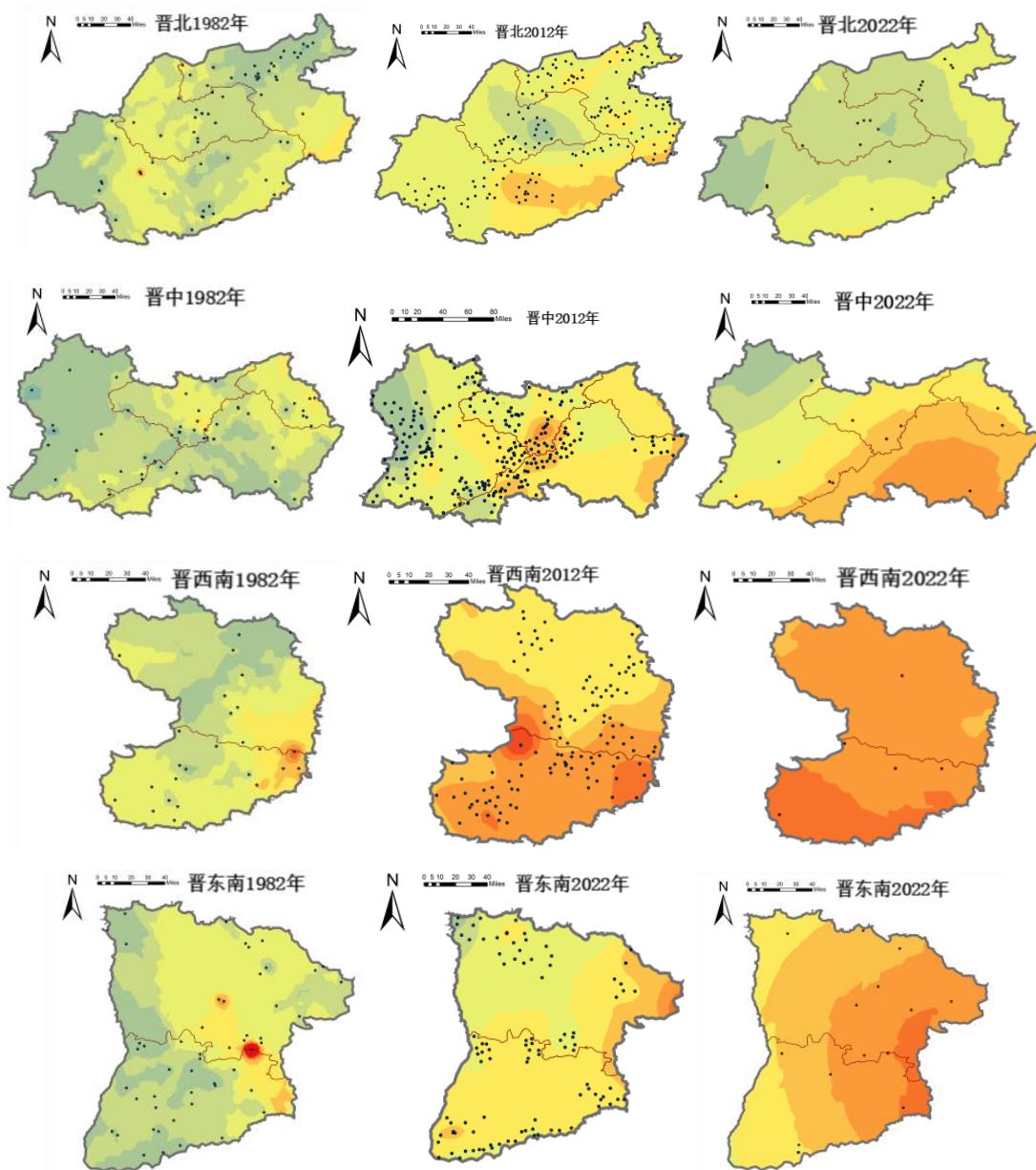


图 3.10 山西省不同区域不同年份全氮含量变化图

### 3.4 本章小结

本章通过分析 1982 年、2012 年、2022 年三年的土壤肥力的指标（有机质、全氮等）的变化和肥力的空间分布，厘清了山西省耕地土壤肥力的变迁。土壤肥力中最重要的养分是有机质和全氮，通过描述性统计分析，可知：（1）有机质含量平均值 1982 年为 9.4g/kg，2012 年为 15.14g/kg，2022 年为 17.59g/kg。有明显

的增长，1982-2012 年增速为 61%，2012-2022 年增速为 16.2%。（2）全氮含量平均值 1982 年为 0.627g/kg，2012 年为 0.853g/kg，2022 年为 0.931g/kg。有明显增长，1982-2012 年增速为 36%，2012-2022 年增速为 9%。

为理解数据地理分布，对有机质和全氮使用克里格插值。二者均呈现对数正态分布，为保证输入模型的数据满足正态分布，对原始数据进行对数变换，变换后的数据基本符合正态分布。下一步对数据拟合半方差函数的理论模型，拟合不同模型寻找平均标准误差最小的模型，计算得到块金值、基台值、偏基台值、块基比等重要数据，得到结论：（1）1982 年有机质含量有中等强度的空间相关性、2012 年和 2022 年有机质含量具有很强烈的空间相关性。（2）1982、2012、2022 年的全氮含量都有很强的空间相关性。

最后，对有机质和全氮含量进行克里格插值。主要结论如下：（1）1982 年山西省有机质含量呈现由西向东增加的趋势，2012 年和 2022 年呈现西偏北向东南增加的趋势。（2）1982 年山西省全氮西侧最低，但整体含量分布较为均匀。2012 年山西省西侧全氮含量没有明显提升，但其他地区出现明显提升，整体上呈现的趋势是西低东南高，其他地区分布较为均匀。2022 年空间分布趋势自西北向东南方向递增。

## 第4章 山西省农田土壤有机质影响因素分析

### 4.1 土壤有机质与养分的关系

如图所示，有机质与全氮、全磷、有效磷、速效钾、缓效钾、碱解氮均有一定程度的正相关关系。尤其是有机质与全氮含极显著正相关。

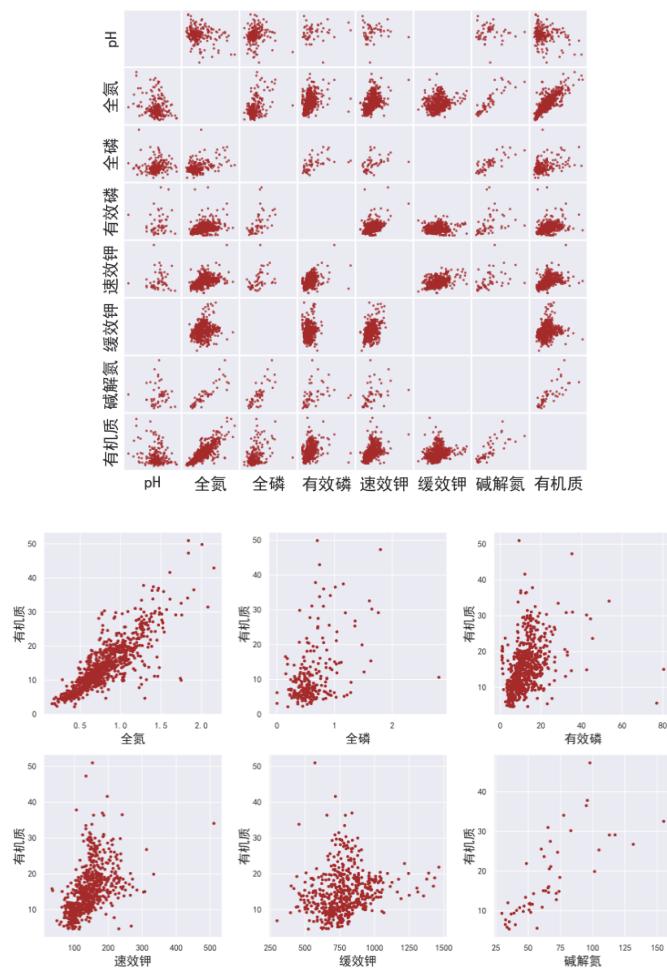


图 4.1 有机质含量及各养分含量间关系散点图

为了更加深入地理解各养分间的相关性关系，进一步计算 Pearson 相关性系数，同时计算 p 值，绘制相关性热力图如下图所示。相关系数约接近 1，说明具有越强烈的线性相关性。此外也需要关注 p 值，p 值小于 0.05 时，认为此次实验

的结果是显著的，在显著的前提下，才有意义参考  $r$  值，进而判断相关程度。

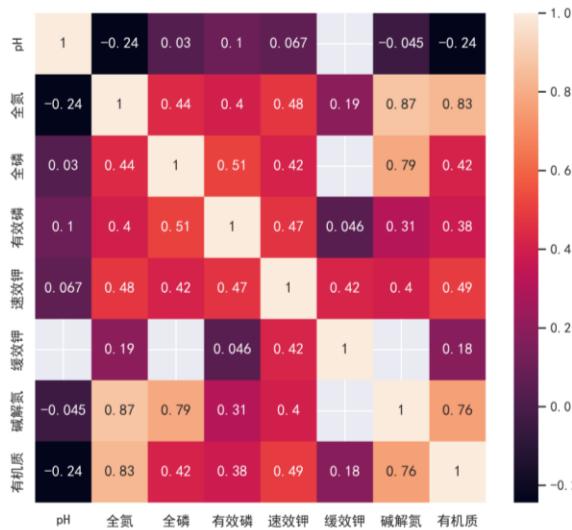


图 4.2 各养分含量相关性热力图

下表是有机质含量和其他养分变量的相关性结果图，此处得到的  $p$  值均为显著，相关系数  $r$  具有参考意义。研究结果显示有机质含量与不同变量之间存在不同的相关强度。有机质含量与 pH 呈现出较为弱的线性负相关，而与其他六个养分变量则呈现出不同程度的正相关。然而，许多线性正相关关系都相对较弱。值得注意的是，有机质含量与全氮含量之间的相关系数达到了 0.834，表明二者之间存在显著且极强的线性正相关关系。这一点也可以从先前的散点图中得以确认。此外，有机质含量与碱解氮含量之间也存在强烈的线性正相关，相关系数为 0.762。

表 4.1 有机质与其他养分含量相关系数结果

变量	相关系数 ( $r$ )	$p$ 值	相关强度
pH	-0.244000	0.000132	弱线性负相关
全磷	0.415000	0.000000	中等强度线性正相关
有效磷	0.378000	0.000000	弱线性正相关
速效钾	0.495000	0.000000	中等强度线性正相关
缓效钾	0.182000	0.000021	极弱线性正相关
碱解氮	0.762000	0.000000	强线性正相关
全氮	0.834000	0.000000	极强线性正相关

## 4.2 土壤有机质与地形、气候、植被的关系

### 4.2.1 各影响因素的时空分布

在研究有机质与 7 个地形、气候、植被因子的关系前，首先大致了解因子的时空分布情况。

第一，地形因子包括高程、坡度、坡向。地形数据在几十年的时间尺度上不会发生显著变化。通过在 Google Earth Engine (GEE) 平台上导出数据并进行绘图分析，确实印证了在几十年间地形因子几乎没有发生变化。因此，本部分只绘制了 2022 年的地形数据。

#### (1) 高程 Elevation

高程表示地面相对于某个参考面的高度，它与海拔高度的变化趋势是一致的，只是参考面不同。通过高程图，可以观察到山西省地形的分布情况。其中，汾河流域的海拔最低，而周围山脉分布区域的海拔较高。

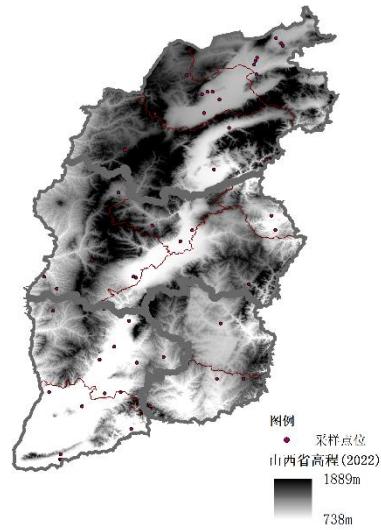


图 4.3 山西省高程图

#### (2) 坡度 Slope

坡度是地面与水平面之间的夹角，用来反映地形的陡峭程度。通过观察，我们可以发现山西省的地形分布具有独特的特色。在汾河中间流域穿过的区域，坡度呈现出连续的  $0^\circ$  条状分布，而其他区域则分布着各类山脉，如西侧的吕梁山脉和东侧的太行山脉。在一些地方，坡度甚至高达近  $72^\circ$ ，显示出了地形的极大陡峭性。

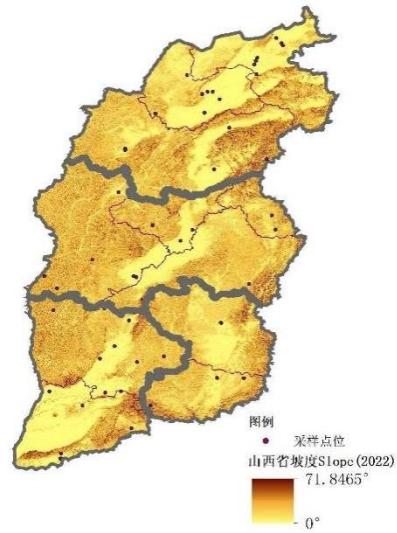


图 4.4 山西省坡度图

### (3) 坡向 Aspect

坡向表示地表面法线与水平面之间的水平角度，能够反映地形的朝向。坡向的取值范围是 0-360 度，0 度代表正北方向，90 度代表正东方向，180 度代表正南，270 度代表正西，靠近 360 度则代表接近正北。坡向分布非常的不规律，唯一较为明显的特征是山西省中间汾河部分的坡向有一条较为明显的向北的坡向。

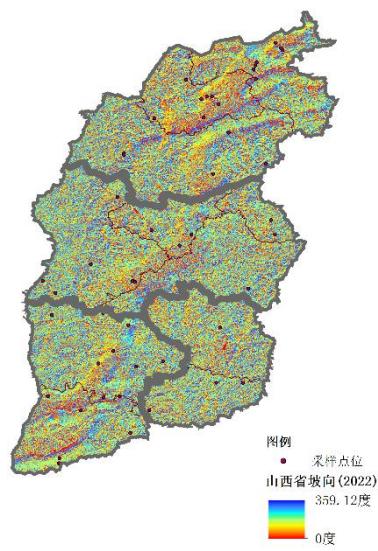


图 4.5 山西省坡向图

第二，气候因子包括年总降水量、年均温、年潜在蒸发量（PET）。

### （1）年总降水量

自 1982-2022 年变换间，年总降水量分布并未发生明显变化。东南部降水量较多，达到 800mm/年，而北部降水量较少，为 300mm/年左右。

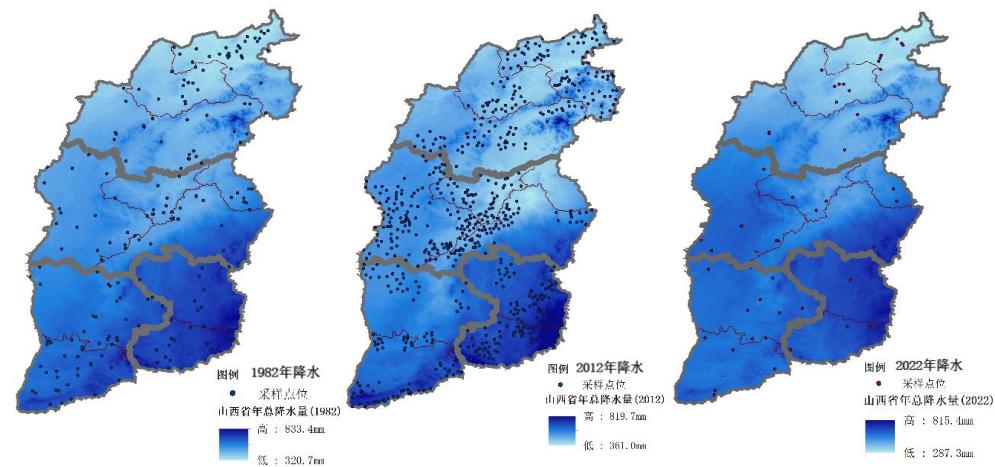


图 4.6 山西省不同年份降水量分布

### （2）年均温

年均温的空间分布在时间上变化不大。它呈现出南高北低的趋势，并与纬度之间存在明显的关系。在山西省内的一些地区，年均温可以达到 15 摄氏度，而在少数地区，年均温则低于零度。

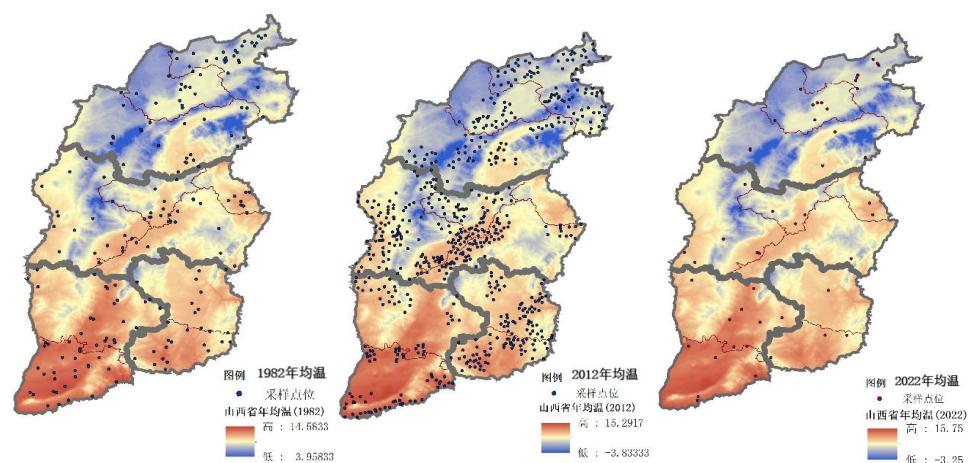


图 4.7 山西省不同年份年均温分布

### (3) 年潜在蒸发量 PET

年潜在蒸发量（PET）的分布在时间上基本没有改变。它呈现出较为不均匀的分布，总体上表现为南高北低的趋势。最高的年 PET 可达到 1300 毫米，而部分地区的 PET 最低约为 450 毫米。

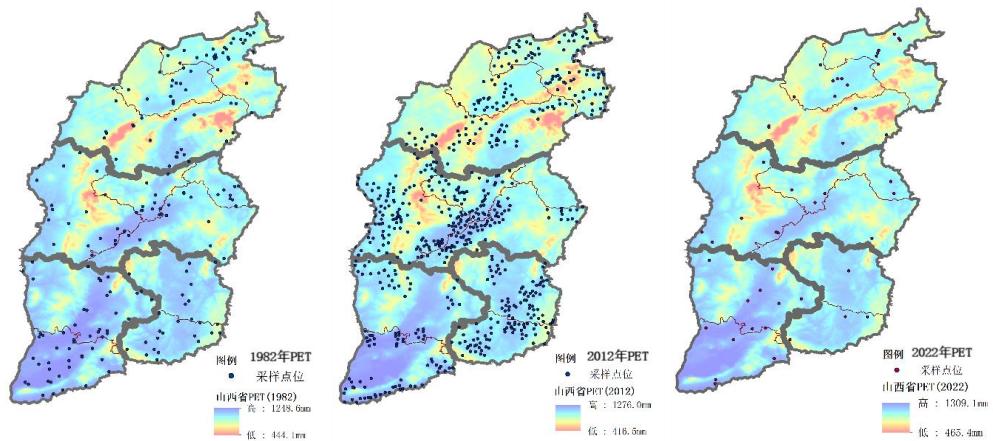


图 4.8 山西省不同年份 PET 分布

第三，植被因子使用归一化植被指数（NDVI）。实际上，植被因子的空间分布在时间上并没有发生很大的变化。NDVI 分布不太均匀，但有几个地区的 NDVI 始终明显高于其他地区。在时间上观察，1982 年的最小 NDVI 值为 362，而 2022 年的最小值为 717，相较于 1982 年翻了一倍。而 1982 年和 2022 年的最大 NDVI 值均约为 5600。可以看出，在 1982 年至 2022 年期间，山西省的植被覆盖有所增加。

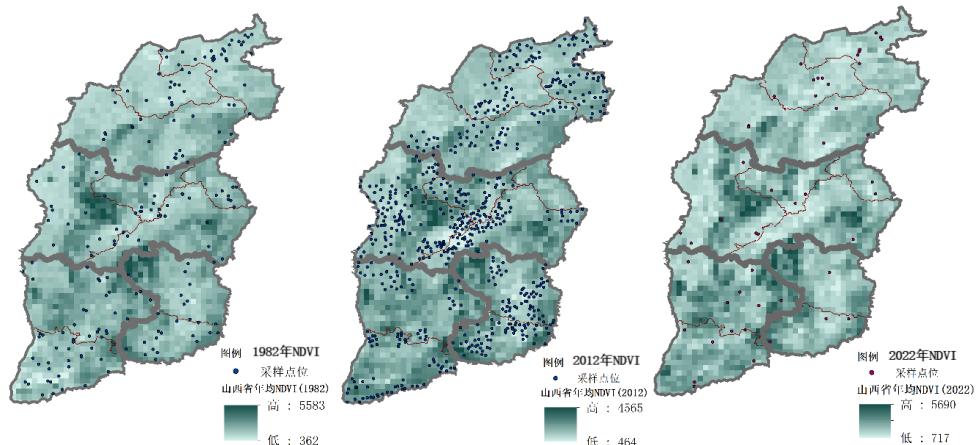


图 4.9 山西省不同年份 NDVI 分布

#### 4.2.2 地形、气候、植被因子对有机质的影响

针对地形、气候和植被因子，绘制了两两之间散点图，并在对角线上绘制各因子的直方图。大部分因子都近似服从正态分布。三对因子之间存在明显的线性关系：高程和潜在蒸发量（PET）、高程和温度、温度和潜在蒸发量。具体而言，随着高程的增加，潜在蒸发量减少；高程的增加也伴随着气温的降低；而温度的升高与潜在蒸发量的增加呈正相关关系。

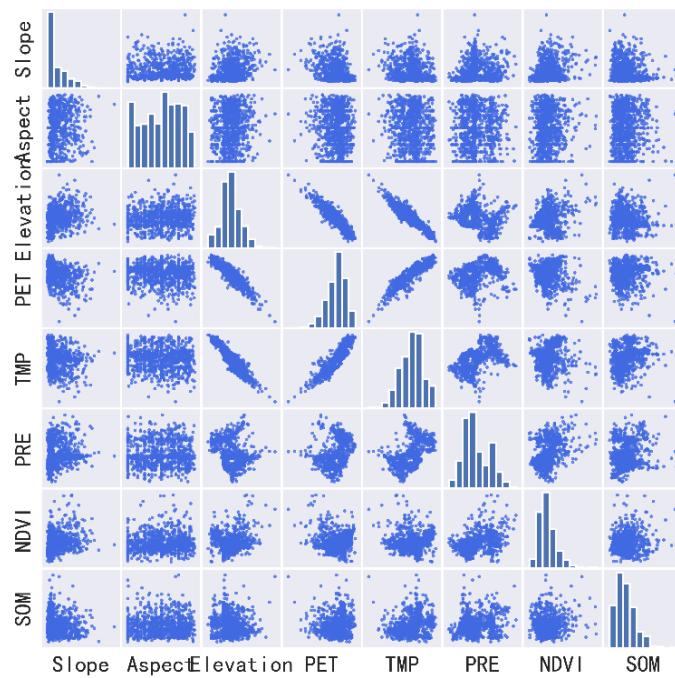


图 4.10 地形、气候、植被因子散点图

将地形的三个因子（坡度、坡向、高程）、气候的三个因子（潜在蒸发量、年均温、降水量）以及植被因子（NDVI）与土壤有机质含量分别绘制了散点图。观察图像，发现这些因子与有机质含量之间的关系并没有明显的线性关系。这表明这些因子对于土壤有机质含量的影响可能是复杂的，并受到其他因素的干扰。进一步的研究和分析可能需要考虑更多的因素和数据。

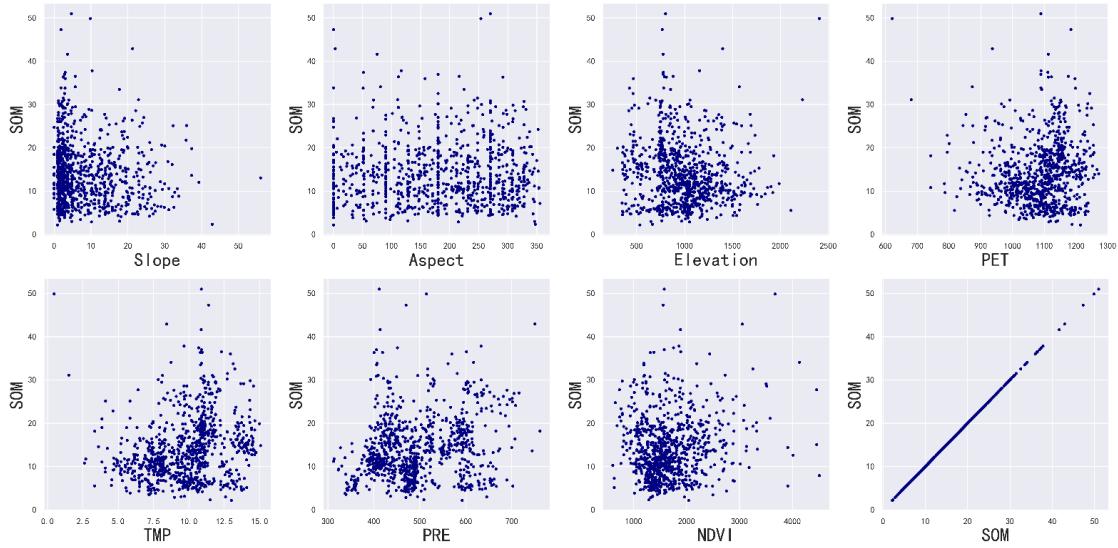


图 4.11 地形、气候、植被因子与有机质含量散点图

同理，进一步计算 Pearson 相关系数，绘制相关性热力图如下图所示。相关系数约接近 1，说明具有越强烈的线性相关性。 $p$  值小于 0.05 时，认为此次实验的结果是显著的。

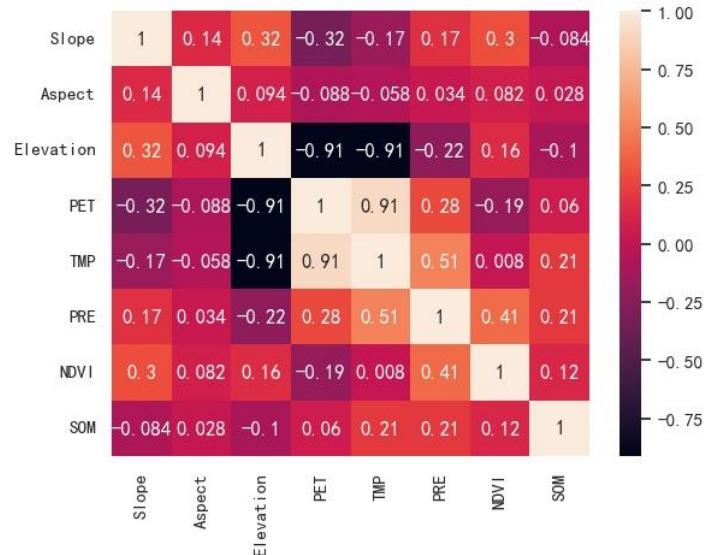


图 4.12 地形、气候、植被因子间相关性热力图

在下表中呈现了有机质含量与其他因子的相关性结果。统计分析发现坡向 (Aspect) 和潜在蒸发量 (PET) 的  $p$  值不具有显著性。然而，其他五个因子的  $p$  值均显著，因此相关系数具有一定的参考意义。有机质含量和这些因子之间并不

存在强烈的线性相关性，其线性相关系数均在 0.3 以下，属于弱相关或极弱相关，甚至可能呈现无相关性。然而，缺乏线性相关性并不意味着这些因子与有机质含量之间没有相关性。本研究将通过第五章的非线性模型分析来进一步探究这些因子与有机质含量之间的关系。

表 4.2 有机质与其他因子相关系数结果（只含 p 为显著的因子）

变量	相关系数 (r)	p 值	相关强度
Slope	-0.084000	0.012875	几乎无线性相关性
Elevation	-0.100000	0.002895	极弱线性负相关/无相关
TMP	0.214000	0.000000	弱线性正相关
PRE	0.212000	0.000000	弱线性正相关
NDVI	0.124000	0.000214	极弱线性正相关/无相关

### 4.3 土壤有机质与农资投入的关系

自 1998 年至 2022 年，山西省耕地灌溉面积发生了改变。1998 年的灌溉面积为 1068.6 千公顷，2007 年增长至 1255.7 千公顷，此后逐步上升，2018 年开始稳定在 1518 千公顷左右，但自 2020 年开始灌溉面积有小幅下降，至 2022 年为 1502 千公顷。

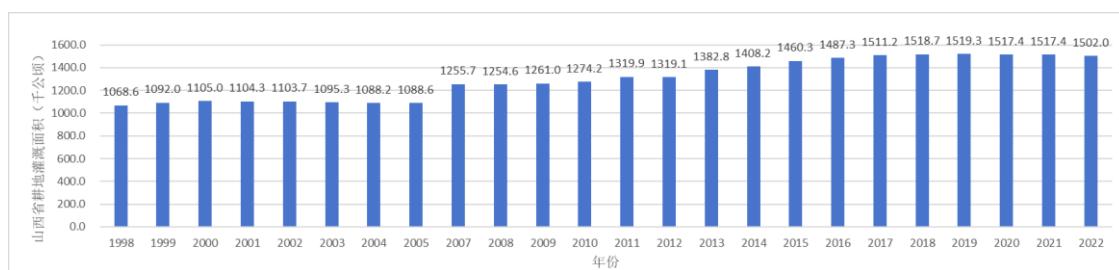


图 4.13 山西省耕地灌溉面积

接下来关注化肥施用量，数据为每千公顷施用的吨化肥。近年来总体化肥施用量呈下降趋势，到 2022 年约为 700 吨化肥/千公顷左右。2012 年化肥施用量达到最高峰，高达 900 吨化肥/千公顷。此外，化肥类型的选择也随着年份发生了显著变化。在 1998 年，氮肥是主要使用的肥料，其次是磷肥和复合肥，而钾肥的使

用量相对较少。然而，复合肥的使用量持续增加，而氮肥和磷肥的使用量大幅下降。到了 2022 年，山西省的总化肥施用量为 689 吨/千公顷，其中复合肥占 463 吨/千公顷的比重，而氮肥占 123 吨/千公顷。这显示出在肥料的选择上，越来越倾向于使用复合肥。

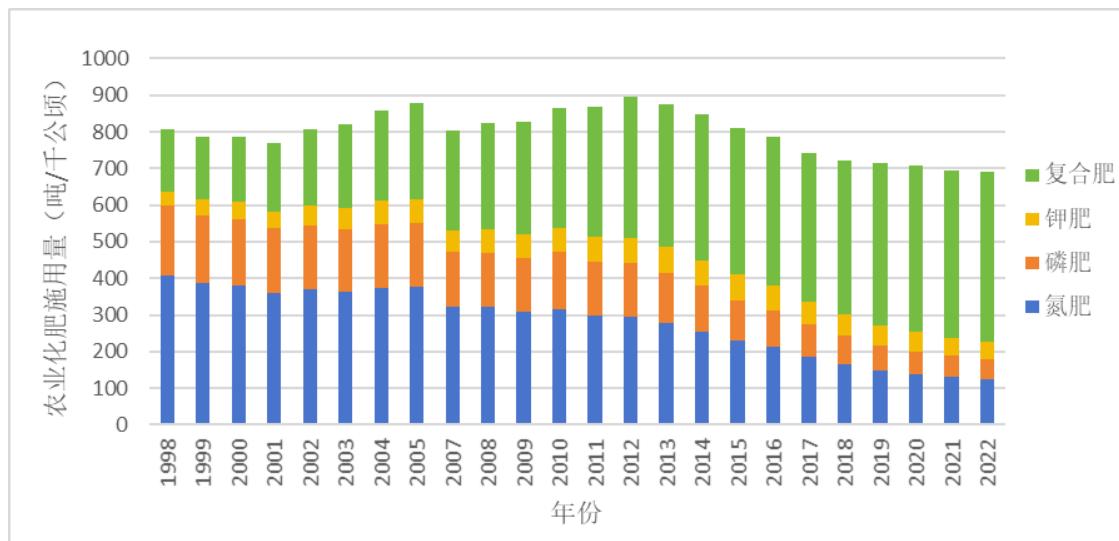


图 4.14 山西省农业化肥施用量

对于山西省整体来说，1982 年的平均有机质含量为 10.28 g/kg，2012 年为 13.91 g/kg，而 2022 年则达到了 29.99 g/kg。同样地，山西省的全氮含量在这段时间内也呈现增长趋势，1982 年的平均含量为 0.72 g/kg，2012 年为 0.78 g/kg，而 2022 年则增至 1.02 g/kg。可见，有机质和全氮含量都呈现出增长态势。

肥料的施用以非常直接的方式向土壤中输送了营养物质，对于土壤养分含量的影响是非常巨大的。化肥的连续施用会显著提高土壤肥力。过往施用的肥料会在长久时间段内对土壤肥力产生影响，多年来积累的肥料施用使得山西省的土壤肥力不断增长。许多文献已经证实肥料对于土壤质量的正向影响，例如周珺<sup>[42]</sup>通过研究长期施肥对番茄土壤有机质组分的影响，通过设计多个实验组和对照组，证实了连续化肥与有机肥配施能够提高土壤肥力。白重九<sup>[43]</sup>通过研究我国北方旱地的有机质变化规律，得出结论：给土壤施肥有机肥，随着施量的增大，土壤有机质含量显著增加。此外，土壤有机质年变化率却呈现先增加后减小的趋势，这也与本研究中山西省的数据相符合。该研究显示当年限超过 25 年，有机质含量的年变化率将会下降。

## 4.4 本章小结

本章分析了各养分、地形、气候、植被、肥料对于有机质含量的影响。（1）通过绘制有机质和各个养分间的散点图、计算 Pearson 相关性系数和 p 值、绘制相关性热力图，发现有机质和 pH 值呈较弱线性负相关，和全氮、碱解氮有强线性正相关，和其他养分均有正线性相关性。（2）通过 ArcMap 绘图，了解了地形、气候、植被因子的时空分布情况。（3）通过绘制地形、气候、植被与有机质含量的散点图以及相关性热力图，发现这些因子与有机质含量的线性相关性都很弱。

（3）通过分析山西省灌溉面积和肥料施用量，了解了山西省农资投入随时间的变化，理解了人为的肥料投入对土壤有机质含量有明显的影响。

## 第5章 山西省农田土壤有机质变化趋势预测

### 5.1 输入模型的因子准备

本章节目的是利用因子预测有机质含量。目前已有的原始因子包含：理化指标、养分因子、气候、地形、植被因子，具体的因子列于下表之中：

表 5.1 原始因子（共 14 个）

类型	因子
理化指标（1 个）	酸碱度（pH）
	全磷（TP）
	有效磷（AP）
	速效钾（AK）
养分（6 个）	缓效钾（SK）
	碱解氮（Alk-N）
	全氮（TN）
	高程（Elevation）
地形（3 个）	坡度（Slope）
	坡向（Aspect）
	年总降水量（Pre）
气候（3 个）	年均温（Tmp）
	年潜在蒸发量（PET）
植被（1 个）	归一化植被指数（NDVI）

为增加输入模型的因子的丰富性，基于以上原始因子构建更多的因子，因子的构建如下表所示。后续基于原始因子与新增因子共 22 个，输入模型。

表 5.2 新增因子（共 8 个）

新增因子（8 个）	构建原因
AP/TP	有效磷与全磷的比例关系
AK/SK	速效钾与缓效钾的比例关系
TN/TP	氮磷比，对于植物生长有实际意义 <sup>[44]</sup>
TN/Alk-N	在第四章分析中发现二者有线性关系
PET/Pre	二者比值为干燥度指数 <sup>[45][46]</sup>
Elevation/PET	在第四章分析中发现二者有线性关系
Tmp/PET	在第四章分析中发现二者有线性关系
Elevation/Tmp	在第四章分析中发现二者有线性关系

## 5.2 模型构建

首先，在将数据输入模型之前要进行缺失值处理，使用因子的平均值填充数据中该因子的缺失值。接着分别训练各个模型，对模型做参数调优和特征筛选。

### 5.2.1 多元线性回归

使用基础的多元线性回归模型对数据进行处理，多元线性回归采用最小二乘法。首先对于数据随机的按照 8: 2 的比例划分为训练集和测试集。对于训练集使用 statsmodels 的 OLS 最小二乘法拟合线性回归模型，得到的系数如下表所示：

表 5.3 多元线性回归系数

	Coef	Std err	t	p
const	-3.5537	27.825	-0.128	0.898
Alk-N	0.0200	0.028	0.724	0.469
Aspect	0.0021	0.001	1.431	0.153
AP	0.0386	0.035	1.109	0.268
Elevation	0.0071	0.008	0.873	0.383
AK	0.0024	0.007	0.355	0.723
TN	6.3534	0.696	23.500	0.000

续表 5.3 多元线性回归系数

	Coef	Std err	t	p
TP	2.7714	1.185	2.340	0.020
Slope	0.0530	0.023	2.277	0.023
SK	0.0035	0.002	2.118	0.035
pH	-1.3522	0.724	-1.867	0.062
PET	0.0186	0.028	-0.669	0.504
TMP	1.1620	2.205	0.527	0.598
PRE	0.0005	0.011	-0.043	0.966
NDVI	0.0006	0.000	-1.712	0.087
AP/TP	0.0095	0.050	0.191	0.849
AK/SK	6.2149	5.647	2.872	0.004
PET/PRE	3.2789	2.511	1.306	0.192
TN/TP	0.8038	0.534	1.506	0.133
TN/Alk-N	238.1226	196.962	1.209	0.227
Elevation/PET	-3.4774	8.359	-0.416	0.678
TMP/PET	8.4780	2296.216	0.004	0.997
Elevation/TMP	0.0039	0.001	2.686	0.007

此时线性回归模型可以写为：

$$\begin{aligned}
 SOM = & -3.5537 + 0.02 * Alk - N + 0.0021 * Aspect + 0.0386 * AP + 0.0071 \\
 & * Elevation + 0.0024 * AK + 16.3534 * TN + 2.7714 * TP - 0.0530 \\
 & * Slope + 0.0035 * SK - 1.3522 * pH - 0.0186 * PET + 1.162 * TMP \\
 & - 0.0005 * PRE - 0.0006 * NDVI + 0.0095 * AP/TP + 16.2149 \\
 & * AK/SK + 3.2789 * PET/PRE + 0.8038 * TN/TP + 238.1226 \\
 & * TN/AlkN - 3.4774 * Elevation/PET + 8.478 * TMP/PET + 0.0039 \\
 & * Elevation/TMP
 \end{aligned} \tag{5-1}$$

在测试集上验证此线性回归模型的结果，得到几个指标如下表所示：

表 5.4 多元线性回归模型表现

模型	R2	MSE	RMSE	MAE
多元线性回归	0.75	11.83	3.44	2.48

值得注意的是并不是每一个自变量都有显著的线性关系，在结果中，有些因子的 t 检验显示其 p 值大于 0.05。因此，需要对变量进行逐步剔除。首先，从该模型中剔除具有最大 p 值且大于 0.05 的因子，并使用剩余的因子构建新模型。然后，从新模型中找出具有最大 p 值且大于 0.05 的因子，如此反复进行，直到最终模型中的所有因子的 p 值均小于 0.05。

表 5.5 经过因子筛选后的多元线性回归模型

	coef	std err	t	p
const	-30.0237	4.345	-6.910	0.000
TN	16.8158	0.545	30.869	0.000
TP	2.9658	0.914	3.243	0.001
Slope	-0.0447	0.019	-2.379	0.018
SK	0.0036	0.001	3.160	0.002
AK/SK	17.3563	3.475	4.994	0.000
PET/PRE	3.1586	0.568	5.561	0.000
TN/TP	0.9284	0.431	2.154	0.032
Elevation/PET	4.2642	1.150	3.706	0.000
TMP/PET	1202.4651	263.291	4.567	0.000
Elevation/TMP	0.0027	0.001	3.148	0.002

模型公式为：

$$\begin{aligned} SOM = & -30.0237 + 16.8158 * TN + 2.9658 * TP - 0.0447 * Slope + 0.0036 * SK \\ & + 17.3563 * AK/SK + 3.1586 * PET/PRE + 0.9284 * TN/TP \\ & + 4.2642 * Elevation/PET + 1202.4651 * TMP/PET + 0.0027 \\ & * Elevation/TMP \end{aligned} \quad (5-2)$$

再在测试集上验证模型效果，结果如下表所示：

表 5.6 多元线性回归模型在因子选择后的表现

模型	R2	MSE	RMSE	MAE
多元线性回归	0.78	10.49	3.24	2.33

通过适当的因子选择，我们可以观察到与直接将 22 个因子作为自变量拟合 Y 相比，仅使用 10 个因子进行拟合的模型表现更加出色。

### 5.2.2 支持向量机回归

构建 SVR 模型，输入所有 22 个因子，使用网格搜索 GridSearchCV 来寻找模型最佳参数，得到最佳组合为：{'C': 2000, 'gamma': 0.0001, 'kernel': 'rbf'}，在测试集上验证这组参数的模型效果，见下表：

表 5.7 SVR 模型在超参数调优后的模型表现

模型	R2	MSE	RMSE	MAE
支持向量机	0.77	11.19	3.35	2.39

在参数调优的过程中，观察到 SVR 模型对参数的影响非常敏感，其性能在参数变动时可能呈现出明显差异。SVR 模型不恰当的参数选择可能导致模型性能显著下降。此外，在模型运行过程中，还发现 SVR 模型相对于其他模型而言的运行时间较长。这是由于求解超平面时需要进行复杂的优化计算以及核函数内积的计算，消耗大量时间。因此，SVR 模型在处理大规模数据的应用场景中可能并不是最合适的选择。

### 5.2.3 随机森林

首先输入所有 22 个因子 X 以及预测目标 Y（即有机质含量），对数据划分为 80% 的训练集和 20% 测试集。使用 RandomForestRegressor 构造随机森林模型，接着使用 GridSearchCV 对于模型中超参数进行寻优，对于 n\_estimators 和 max\_depth 两个超参数进行 5 折交叉验证网格寻优。此步之后得到得分最高的超参数，即 {'max\_depth': 15, 'n\_estimators': 150}，作为随机森林模型中设置的超参数。经过超参数调优后的随机森林模型在测试集上得到的模型表现如下表所示。

表 5.8 随机森林模型在超参数调优后的模型表现

模型	R2	MSE	RMSE	MAE
随机森林	0.81	8.77	2.96	1.95

对于输入模型的因子，获得因子重要性，根据降序绘制特征重要性排序图：

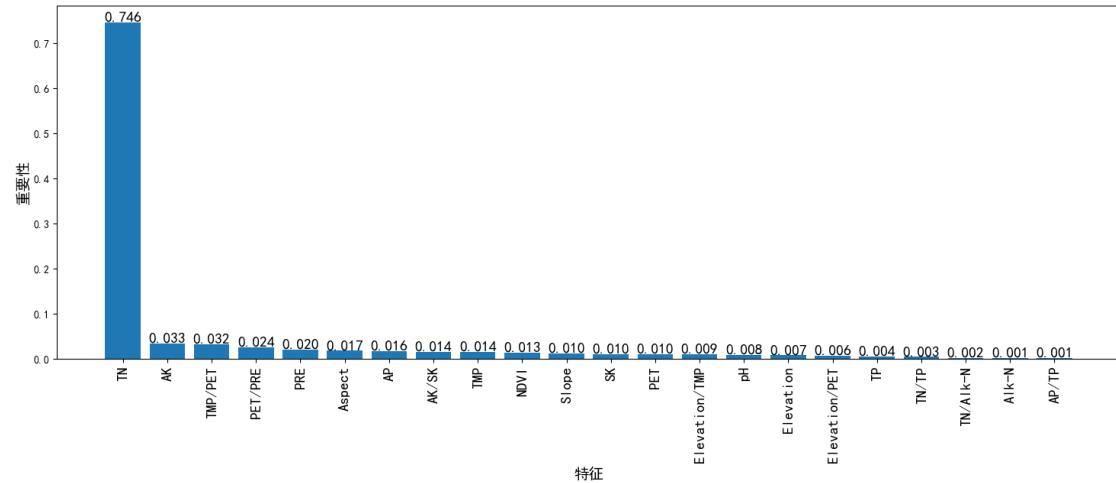


图 5.1 随机森林特征重要性

观察特征重要性后发现，在随机森林模型中，全氮是对于有机质含量预测起到非常重要作用的因子。其次是全钾、年均温与年均潜在蒸发量的比值。然而，还有一些因子被认为不太重要，可以在特征筛选中进行剔除。通过尝试不同数量的特征剔除，发现将重要性阈值定为 0.05 时，模型的表现效果最佳。因此，剔除了 5 个特征，包括全磷、全氮与全磷的比值、全氮与碱解氮的比值、碱解氮以及有效磷与全磷的比值。值得注意的是，被剔除的特征主要涉及氮和磷相关的特征。这可能是因为在最重要的特征中已经包含了这两种元素，它们提供了相关的信息。

使用剩余的 17 个特征和调优后的超参数构建随机森林模型，同样在测试集上进行验证，经过特征优化后的模型表现如下表所示：

表 5.9 随机森林模型在超参数调优及特征优化后的模型表现

模型	R2	MSE	RMSE	MAE
随机森林	0.81	8.62	2.94	1.94

通过超参数调优和特征优化后，观察到模型的表现明显优于仅进行超参数调优的情况。根据评估指标 MSE、MAE 和 RMSE 的下降情况来看，特征优化很可能剔除了一些冗余信息，从而提升了模型的性能。

#### 5.2.4 XGBoost

与随机森林的过程类似，经过前序操作后进行超参数优化，设置网格寻优的超参数范围如下表所示：

表 5.10 XGBoost 超参数寻优范围

超参数	寻优范围
Max_depth	[3, 5, 7, 9]
Min_child_weight	[1, 3, 5, 7]
N_estimators	[50,100, 150]
Learning_rate	[0.01, 0.02, 0.05, 0.1]
gamma	[0, 0.05, 0.3, 0.5, 0.7, 0.9, 1]

最终得到的最优超参数组合为：{'gamma': 0.3, 'learning\_rate': 0.1, 'max\_depth': 5, 'min\_child\_weight': 3, 'n\_estimators': 150}。此模型的表现如下表所示：

表 5.11 XGBoost 模型在超参数调优后的模型表现

模型	R2	MSE	RMSE	MAE
XGBoost	0.83	7.74	2.78	1.96

对于输入模型的因子，获得因子重要性，根据降序绘制特征重要性排序图：

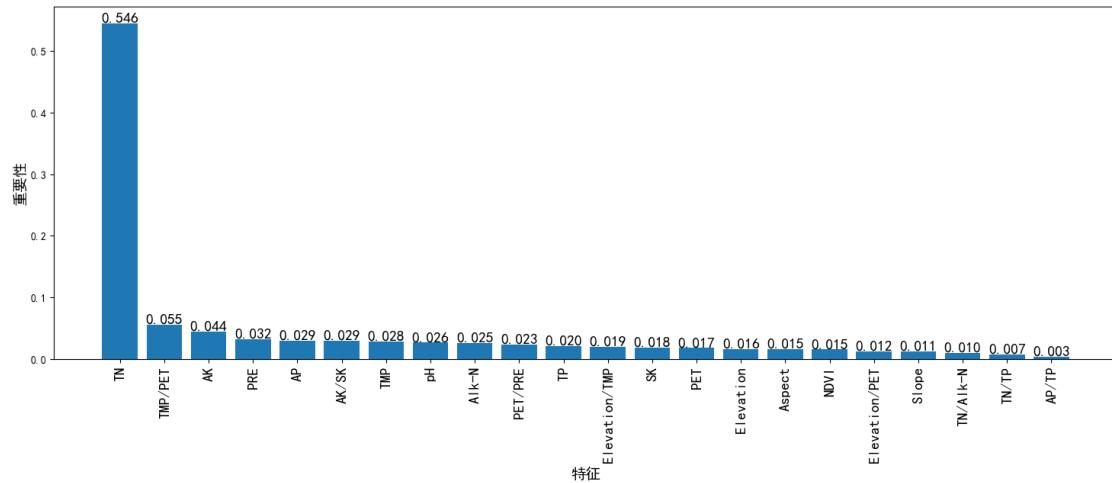


图 5.2 XGBoost 特征重要性

观察特征重要性后发现，在 XGBoost 模型中，仍然是全氮对于有机质含量预测起到非常重要的作用。其次是年均温与年均潜在蒸发量的比值、全钾和降水。总体而言，特征的重要程度与随机森林模型相似，尤其是最重要的几个特征。然而，与随机森林相比，XGBoost 模型的特征重要性分布相对更均衡一些，全氮的相对重要性有所下降。针对特征剔除，设定不同的阈值，分别尝试剔除重要性低于 0.01、0.015、0.017、0.02 的特征，并验证模型效果。结果显示，剔除重要性低于 0.015 的特征可以获得最佳的模型效果。因此，剔除了 5 个特征，包括有效磷/全磷、全氮/全磷、全氮/碱解氮、坡度以及高程/潜在蒸发量。使用剩余的 17 个特征和经过调优的超参数构建 XGBoost 模型，并在测试集上进行验证。特征优化后的模型表现如下表所示：

表 5.12 XGBoost 模型在超参数调优及特征优化后的模型表现

模型	R2	MSE	RMSE	MAE
XGBoost	0.84	7.29	2.70	1.92

超参数调优和特征优化后的模型表现要优于仅仅是超参数调优的模型表现。指标 mse、mae 和 rmse 都有所下降。

### 5.2.5 LightGBM

LightGBM 仍然是基于树的模型，同理进行超参数优化，设置网格寻优的超参数范围如下表所示：

表 5.13 LightGBM 超参数寻优范围

超参数	寻优范围
Max_depth	[5,10,15]
Learning_rate	[0.02, 0.05, 0.1]
Feature_fraction	[0.6, 0.8, 0.95]
Bagging_fraction	[0.6, 0.8, 0.9]

最优超参数组合为：{'bagging\_fraction': 0.6, 'feature\_fraction': 0.95, 'learning\_rate': 0.1, 'max\_depth': 15}。此模型的表现如下表所示：

表 5.14 LightGBM 模型在超参数调优后的模型表现

模型	R2	MSE	RMSE	MAE
LightGBM	0.83	7.62	2.76	1.93

对于输入模型的因子，获得因子重要性，根据降序绘制特征重要性排序图：

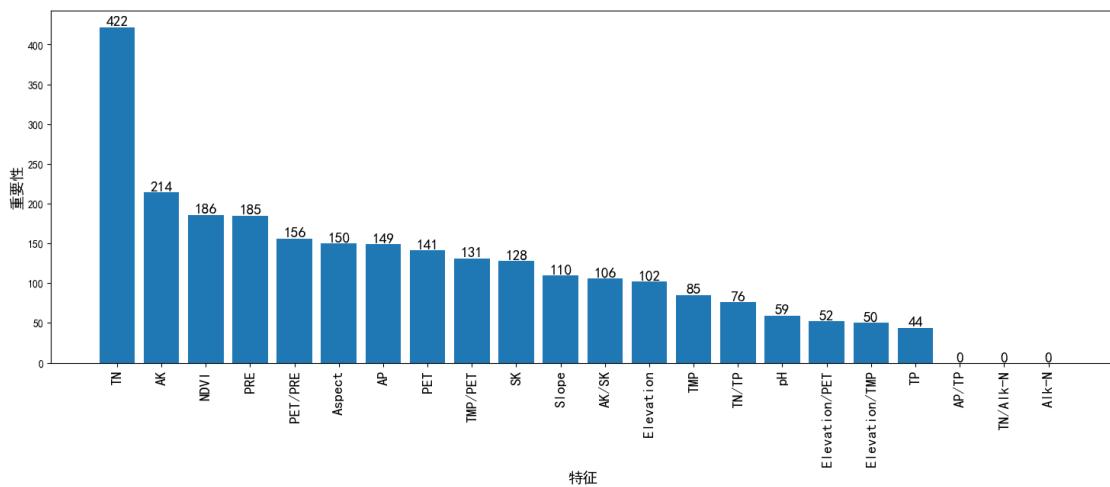


图 5.3 LightGBM 特征重要性

经过观察发现，LightGBM 模型的特征重要性分布相比前两个树模型更加均衡，这表明该模型能够更好地处理和利用不同的特征。虽然全氮仍然排在第一位，但后面几个特征的排序与随机森林和 XGBoost 有所不同，包括全钾、NDVI、降水、潜在蒸发量与降水量的比值。相对于前两个模型，气候条件在 LightGBM 模型中起到了更为重要的作用。在特征剔除方面，设置了不同的阈值并验证了模型效果。结果显示，剔除最不重要的 7 个特征可以获得最佳的模型效果。因此，我们剔除了碱解氮、全氮/碱解氮、有效磷/全磷、全磷、高程/年均温、高程/潜在蒸发量以及 pH 这 7 个特征。使用剩余的 15 个特征和经过调优的超参数构建 LightGBM 模型，并在测试集上进行验证。特征优化后的模型表现如下表所示：

表 5.15 LightGBM 模型在超参数调优及特征优化后的模型表现

模型	R2	MSE	RMSE	MAE
LightGBM	0.84	7.47	2.73	1.85

结果显示，超参数调优和特征优化的组合策略对于改善模型性能具有重要意义。

### 5.3 模型效果比较

上一节中分别使用多元线性回归、支持向量机回归、随机森林、XGBoost 和 LightGBM 构建了有机质含量的预测模型。对各个模型都进行了不同程度的参数调优和特征选择。将模型在测试集上的表现整理于下表中：

表 5.16 模型效果比较

模型	是否 参数调优	是否 筛选特征	特征数	R2	MSE	RMSE	MAE
多元线性回归	/	否	22	0.75	11.83	3.44	2.48
	/	是	10	0.78	10.49	3.24	2.33
支持向量机	是	否	22	0.77	11.19	3.35	2.39
	是	否	22	0.81	8.77	2.96	1.95
随机森林	是	是	17	0.81	8.62	2.94	1.94

续表 5.16 模型效果比较

模型	是否参数调优	是否筛选特征	特征数	R2	MSE	RMSE	MAE
XGBoost	是	否	22	0.83	7.74	2.78	1.96
	是	是	17	0.84	7.29	2.70	1.92
LightGBM	是	否	22	0.83	7.62	2.76	1.93
	是	是	15	0.84	7.47	2.73	1.85

上述表格，可以明确显示到上一节中构建的所有模型的表现。首先，通过参数调优和特征筛选，同类型的模型都得到了改善。参数调优使得模型能够找到最适合当前场景的超参数，而特征筛选则有助于去除冗余和复杂信息。其次，对经过调优的不同模型进行比较，可以得出以下排序关系：第一，根据 R2 指标排序： $\text{LightGBM}=\text{XGBoost}>\text{随机森林}>\text{多元线性回归}>\text{支持向量机}$ 。第二，根据 MSE 和 RMSE 指标排序： $\text{XGBoost}>\text{LightGBM}>\text{随机森林}>\text{多元线性回归}>\text{支持向量机}$ 。第三，根据 MAE 指标排序： $\text{LightGBM}>\text{XGBoost}>\text{随机森林}>\text{多元线性回归}>\text{支持向量机}$ 。

综合考虑各个模型表现的指标，可以得出结论：在本研究的场景下，LightGBM 和 XGBoost 表现最佳，不仅在预测准确性方面优秀，而且运行效率也相对较高。其次是随机森林，而线性回归和支持向量机的表现相对较差，与决策树类模型相比差距较大。

## 5.4 设定未来场景预测

基于 5.3 对于各个模型的分析，最终选定使用 LightGBM 和 XGBoost 模型对于未来场景进行预测。对未来场景进行设计，是指通过预设表中原始指标：理化指标、养分指标、地形指标、气候指标、植被指标的未来变化，并用此预设情形来预测该情景下有机质的含量。

首先建立如今的现状数据，以经过筛选后 2022 年原始指标的平均值，作为现状数据，如下表所示：

表 5.17 指标现状

类型	因子	现状值
预测目标	SOM	17.58981
理化指标（1个）	酸碱度 (pH)	8.322404
	全磷 (TP)	0.814308
	有效磷 (AP)	20.40335
	速效钾 (AK)	158.532
养分（6个）	缓效钾 (SK)	778.7418
	碱解氮 (Alk-N)	64.44679
	全氮 (TN)	0.930731
	高程 (Elevation)	953.4038
地形（3个）	坡度 (Slope)	4.869538
	坡向 (Aspect)	171.1872
	年总降水量 (Pre)	483.6654
气候（3个）	年均温 (Tmp)	10.4234
	年潜在蒸发量 (PET)	1136.456
植被（1个）	归一化植被指数 (NDVI)	2103.846

接下来假设不同的未来情景。首先，由数据上的变化趋势理解因子未来可能的变动情况。第一，由第三章中描述性统计可知 pH 值自 1982 年至 2022 年的四十年间增长了 2.46%。第二，对于养分指标，TP 四十年间相对增长 45%，AP 四十年间相对增长 76%，AK 在近 10 年间增长了 15%。SK 和 Alk-N 由于只有一年的数据，并且微量元素变化不大，近似假定为不会发生变化。TN 2012-2022 年增长率为 9%。第三，由于地形的常年不变性，始终假定三个地形指标不变化。第四，对于气候指标，查找文献<sup>[47]</sup>后得到山西省近 60 年间的平均气温增长率为 0.396 度/10a；降水量递减，递减率为 3.03mm/10a。同理查找文献<sup>[48]</sup>后得到山西省近 50 年的年平均潜在蒸发量递减，递减率为 0.3mm/10a。第五，对于 NDVI，文献<sup>[49]</sup>中显示，自 2005-2015 年十年间增长率为 0.0106/年。

其次，考虑人为政策和耕作措施的改变，在可持续发展和 3060 的理念之下，山西省对于植树造林的重视程度增强，同时进行产业转型，恢复绿化，归一化植

被指数增加，年均温增加趋势放缓，年总降水量减少趋势放缓，PET 减少趋势放缓。此外对于土壤肥力的重视和肥料的投入使得土壤中养分含量出现不同程度的增长。

综合以上两层考虑方向，将假设情景定位为 10 年后，可定义较为可能的变化情形：

表 5.18 变化情形

类型	因子	10 年后相比现在变化
理化指标（1 个）	酸碱度 (pH)	不变
	全磷 (TP)	1.1125 倍
	有效磷 (AP)	1.19 倍
	速效钾 (AK)	1.15 倍
	缓效钾 (SK)	不变
	碱解氮 (Alk-N)	不变
	全氮 (TN)	1.09 倍
地形（3 个）	高程 (Elevation)	不变
	坡度 (Slope)	不变
	坡向 (Aspect)	不变
气候（3 个）	年总降水量 (Pre)	-3.03mm
	年均温 (Tmp)	+0.396°C
	年潜在蒸发量 (PET)	-0.3mm
植被（1 个）	归一化植被指数 (NDVI)	+0.106

使用训练好的 XGBoost 模型预测此假定情景下 10 年后的有机质含量，得到预测结果为 20.004787g/kg。同理，使用训练好的 LightGBM 模型对该情形进行预测，10 年后有机质含量的预测结果为 19.60664391g/kg。二者的相对误差仅为 2.01%，说明 XGBoost 和 LightGBM 模型对于这个情景下的有机质含量预测的结果是非常接近的，二者的相近也印证了预测结果的可靠性。将结果总结与下表之中：

表 5.19 XGBoost 与 LightGBM 模型预测结果

模型	十年后有机质含量预测结果	当下有机质含量	增长率
XGBoost	20.004787	17.58981	13.73%
LightGBM	19.60664391	17.58981	11.47%

## 5.5 本章小结

本章从因子准备、模型构建、模型效果比较和预测未来假定情形，使用理化指标、养分、地形、气候、植被指标预测有机质含量。第一，基于 14 个原始因子构建了 8 个新增因子，共计 22 个因子输入模型。第二，构建了 5 个预测模型，并比较了结果。使用多元线性回归模型并通过 t 检验筛选了因子，得到了拟合公式；使用支持向量机模型并对参数调优，得到了模型表现；使用随机森林、XGBoost 和 LightGBM 模型，进行调优并通过特征重要性分布筛选特征。对于模型均发现超参数调优和特征选择能够有效改进模型效果。第三，通过对比模型在测试集上的表现，得出 LightGBM 和 XGBoost 表现最佳的结论。第四，通过分析数据趋势，并结合人为政策和耕作措施，设定 10 年后各因子指标的变化情形。使用训练好的 XGBoost 和 LightGBM 模型预测 10 年后情形下有机质含量，两模型得出极为接近的结果，分别预测有机质含量为 20.0g/kg 和 19.6g/kg，相应计算得到有机质含量的增长率为 13.73% 和 11.47%。

## 第 6 章 结论与展望

### 6.1 结论

本文研究了山西省农田土壤有机质含量和其他土壤养分含量的演变特征，理解了有机质含量背后的影响因素，最终建立并优化了机器学习模型，并且假定了未来情境并给出了该情景下有机质含量的预测值。主要结论有以下几点：

第一，山西省 1982 年有机质含量为 9.4g/kg，2012 年为 15.14g/kg，2022 年为 17.59g/kg，逐年递增。1982 年全氮含量为 0.627g/kg，2012 年为 0.853g/kg，2022 年为 0.931g/kg，1982-2012 年增速为 36%，2012-2022 年增速为 9%。通过计算半方差函数，拟合理论模型，并完成克里格插值得到以下结论：1982 年有机质含量为中等强度的空间相关性，2012 年、2022 年具有很强空间相关性。从空间上分布趋势来说，1982 年有机质含量由西向东增加，2012 年及 2022 年呈西北向东南增加态势。

第二，本文通过皮尔森相关系数及相关性热力图，得出结论：有机质和 pH 值呈较弱线性负相关，和全氮、碱解氮有强线性正相关，和其他养分均有正线性相关性。然而地形、气候、植被因子与有机质含量的线性关系并不明显。此外，通过分析山西省灌溉及肥料施用的农资投入，得出结论：人为肥料输入对有机质含量有明显影响。

第三，本文通过特征构建、超参数优化和特征筛选，建立了理化指标、养分、地形、气候、植被指标与有机质含量之间的关系。使用简单模型如多元线性回归，并使用支持向量机回归以及三类决策树相关模型（随机森林、XGBoost、LightGBM），分别搭建了模型预测有机质含量。经过参数调优和特征筛选均得到了更为优化的结果。最终通过验证模型在测试集上的效果，得出结论：LightGBM 和 XGBoost 在此预测情景下表现最佳。

第四，通过分析数据趋势，以及考量人为政策及工作措施，合理设定 10 年后各理化指标、养分、地形、气候、植被的变化情况，利用训练好的 XGBoost 和 LightGBMt 模型预测未来该情境下有机质含量。二者得出相近结果：分别预测得到未来该场景下有机质为 20.0g/kg 和 19.6g/kg，也即相应的增长率为 13.73% 和 11.47%。

总而言之，本研究通过了解现象、理解影响因素、预测未来的三个步骤，逐步深入了解了山西省农田土壤有机质的变化情况和背后的作用机制，并且尝试对未来特定场景下的含量进行了预测。

## 6.2 不足与展望

本研究仍然存在一定的不足和局限性。

第一，本文数据量并不充裕。数据主要来源于文献资料和实地调研，样本总量相对较少，这导致了在构建模型进行预测时精度可能有所下降。未来的研究中考虑多进行现场实测采样并记录数据，通过实地采样获得更多可靠数据。

第二，本文对人为影响的分析不够充分。由于化肥数据难以收集，仅有山西省作为一个整体的肥料施用数据，因此对于农资影响的分析只能停留在较为宏观的层面。未来考虑收集更多化肥数据，并且通过寻找间接数据推测农资投入。

第三，本文对于未来十年后各因素的情形假定可以优化。研究中仅考虑了数据本身趋势及定性地考虑了人为的耕作习惯的改变，但是并没有依托有说服力的相关研究。未来的研究中可以考虑广泛搜寻学者们对未来气候（如降水、均温、潜在蒸发量）以及植被情况的预测，依托于这部分有说服力的预测结果构建未来情景。

综上，本文还有许多的改进空间，会在未来的研究中积极考虑并加以优化。

## 插图索引

图 1.1 大气二氧化碳增速 .....	1
图 1.2 全球碳汇 .....	2
图 1.3 技术路线 .....	8
图 2.1 山西省行政区域划分 .....	9
图 2.2 不同年份土样采集点位分布 .....	10
图 3.1 1982 有机质数据分布 .....	20
图 3.2 1982 年变换前（左）后（右）有机质数据正态 QQ 图.....	20
图 3.3 1982 年山西省耕地有机质半方差函数 .....	21
图 3.4 2012 年变换后有机质数据正态 QQ 图 .....	22
图 3.5 2012 年山西省耕地有机质半方差函数 .....	23
图 3.6 2022 年山西省耕地有机质半方差函数 .....	24
图 3.7 山西省不同年份农田土壤有机质含量分布 .....	24
图 3.8 山西省不同区域不同年份有机质含量变化图 .....	26
图 3.9 山西省农田土壤不同年份全氮含量变化 .....	28
图 3.10 山西省不同区域不同年份全氮含量变化图 .....	29
图 4.1 有机质含量及各养分含量间关系散点图 .....	31
图 4.2 各养分含量相关性热力图 .....	32
图 4.3 山西省高程图 .....	33
图 4.4 山西省坡度图 .....	34
图 4.5 山西省坡向图 .....	34
图 4.6 山西省不同年份降水量分布 .....	35

图 4.7 山西省不同年份年均温分布 .....	35
图 4.8 山西省不同年份 PET 分布 .....	36
图 4.9 山西省不同年份 NDVI 分布 .....	36
图 4.10 地形、气候、植被因子散点图 .....	37
图 4.12 地形、气候、植被因子间相关性热力图 .....	38
图 4.13 山西省耕地灌溉面积 .....	39
图 4.14 山西省农业化肥施用量 .....	40
图 5.1 随机森林特征重要性 .....	47
图 5.2 XGBoost 特征重要性 .....	49
图 5.3 LightGBM 特征重要性 .....	50

## 表格索引

表 2.1 山西省行政区域划分及耕地面积与样点概况 .....	10
表 3.1 1982-2022 年山西省各市农田土壤有机质含量（单位： g/kg） .....	17
表 3.2 土壤肥力指标时间演变 .....	18
表 3.3 1982 年有机质描述性统计 .....	20
表 3.4 1982 有机质不同半方差函数拟合结果 .....	21
表 3.5 2012 年有机质描述性统计 .....	22
表 3.6 2012 有机质不同半方差函数拟合结果 .....	22
表 3.7 2022 年有机质描述性统计 .....	23
表 3.7 2022 有机质不同半方差函数拟合结果 .....	23
表 3.8 山西省耕地全氮含量描述性统计 .....	27
表 3.9 山西省耕地三年全氮克里格插值结果 .....	27
表 4.1 有机质与其他养分含量相关系数结果 .....	32
表 4.2 有机质与其他因子相关系数结果（只含 p 为显著的因子） .....	39
表 5.1 原始因子（共 14 个） .....	42
表 5.2 新增因子（共 8 个） .....	43
表 5.3 多元线性回归系数 .....	43
表 5.4 多元线性回归模型表现 .....	45
表 5.5 经过因子筛选后的多元线性回归模型 .....	45
表 5.6 多元线性回归模型在因子选择后的表现 .....	46
表 5.7 SVR 模型在超参数调优后的模型表现 .....	46
表 5.8 随机森林模型在超参数调优后的模型表现 .....	47

表 5.9 随机森林模型在超参数调优及特征优化后的模型表现 .....	48
表 5.10 XGBoost 超参数寻优范围 .....	48
表 5.11 XGBoost 模型在超参数调优后的模型表现 .....	48
表 5.12 XGBoost 模型在超参数调优及特征优化后的模型表现 .....	49
表 5.13 LightGBM 超参数寻优范围 .....	50
表 5.14 LightGBM 模型在超参数调优后的模型表现 .....	50
表 5.15 LightGBM 模型在超参数调优及特征优化后的模型表现 .....	51
表 5.16 模型效果比较 .....	51
表 5.17 指标现状 .....	53
表 5.18 变化情形 .....	54
表 5.19 XGBoost 与 LightGBM 模型预测结果 .....	55

## 参考文献

- [1] Chen L, Msigwa G, Yang M, Osman AI, Fawzy S, Rooney DW, Yap P-S (2022) Strategies to achieve a carbon neutral society: a review. Environ Chem Lett 20:2277–2310. <https://doi.org/10.1007/s10311-022-01435-8>
- [2] Carbon Dioxide Daily (2022) Latest daily CO<sub>2</sub>. <https://www.co2.earth/daily-co2>. Accessed 28 Nov 2023
- [3] 王灿, 张雅欣.碳中和愿景的实现路径与政策体系[J].中国环境管理, 2020, 12(06):58-64.DOI:10.16868/j.cnki.1674-6252.2020.06.058.
- [4] Carbon Dioxide Acceleration. <https://www.co2.earth/co2-acceleration>. Accessed 28 November 2023
- [5] R. Lal, Soil carbon sequestration to mitigate climate change, Geoderma, Volume 123, Issues 1–2, 2004, Pages 1-22, ISSN 0016-7061, <https://doi.org/10.1016/j.geoderm.a.2004.01.032>.
- [6] Lehmann, J., Kleber, M. The contentious nature of soil organic matter. Nature 528, 60–68 (2015). <https://doi.org/10.1038/nature16069>
- [7] R. Lal, Soil carbon sequestration to mitigate climate change, Geoderma, Volume 123, Issues 1–2, 2004, Pages 1-22, ISSN 0016-7061, <https://doi.org/10.1016/j.geoderm.a.2004.01.032>.
- [8] 程琨, 潘根兴.农业与碳中和[J].科学, 2021, 73(06):8-12+4.
- [9] Lal R. Soil carbon sequestration in China through agricultural intensification, and restoration of degraded ecosystemsc. Land Degradation & Developmen, 2002, 13: 469-478.
- [10] 李娟, 谢冬梅, 祁生源.论土壤有机质含量对提高土壤肥力保护耕地的重要性[J].青海农技推广, 1999(01):57-58.
- [11] 李铮, 王英武.山西省耕地土壤有机质评述[J].山西农业科学, 1992, (11):3-5.
- [12] 朱静, 黄标, 孙维侠等.长江三角洲典型地区农田土壤有机质的时空变异特征及其影响因素[J].土壤, 2006(02):158-165.
- [13] 解文艳, 周怀平, 杨振兴等.山西省农田土壤肥力现状及近 10 年变化特征[J].中国土壤与肥料, 2022, (10):1-10.
- [14] 胡克林, 余艳, 张凤荣等.北京郊区土壤有机质含量的时空变异及其影响因素[J].中国农业科学, 2006(04):764-771.

- [15] 于婧文, 周怀平, 解文艳等.土壤养分变化及农户养分管理现状研究[J].山西农业科学, 2010, 38(06):33-36.
- [16] 林小丁, 窦春宇, 张彩云等.陕西省关中地区耕地土壤属性变化趋势研究[J].植物营养与肥料学报, 2023, 29(10):1853-1862.
- [17] 张建杰, 张强, 杨治平等.山西临汾盆地土壤有机质和全氮的空间变异特征及其影响因素[J].土壤通报, 2010, 41(04):839-844.DOI:10.19336/j.cnki.trtb.2010.04.016.
- [18] 宋莎, 李廷轩, 王永东等.县域农田土壤有机质空间变异及其影响因素分析[J].土壤, 2011, 43(01):44-49.DOI:10.13758/j.cnki.tr.2011.01.006.
- [19] 申若禹, 张吴平, 王国芳等.基于地统计学的山西省不同类型土壤有机质空间变异分析研究[J].山东农业科学, 2019, 51(04):110-116.DOI:10.14083/j.issn.1001-4942.2019.04.021.
- [20] 王国芳, 张吴平, 毕如田等.县域尺度农田深层土壤有机质的估算及空间变异特征[J].农业工程学报, 2019, 35(22):122-131.
- [21] 朱静, 黄标, 孙维侠等.长江三角洲典型地区农田土壤有机质的时空变异特征及其影响因素[J].土壤, 2006(02):158-165
- [22] 胡克林, 余艳, 张凤荣等.北京郊区土壤有机质含量的时空变异及其影响因素[J].中国农业科学, 2006(04):764-771.
- [23] 李浩. 土壤属性时空地统计建模与分析[D].华中农业大学, 2023.DOI:10.27158/d.cnki.ghznu.2022.000003.
- [24] 赵明松, 张甘霖, 李德成等.江苏省土壤有机质变异及其主要影响因素[J].生态学报, 2013, 33(16):5058-5066.
- [25] 李浩. 土壤属性时空地统计建模与分析[D].华中农业大学, 2023.DOI:10.27158/d.cnki.ghznu.2022.000003.
- [26] 黄婷. 基于支持向量机的土壤基础肥力评价和土壤有机质含量预测研究[D].南京农业大学, 2017.
- [27] 王茵茵. 基于 RS 数据与 RF 算法的陕西省土壤有机质预测研究[D].西北农林科技大学, 2016.
- [28] 胡贵贵. 基于环境变量的苹果区土壤养分空间预测及其影响因子分析[D].西北大学, 2021.DOI:10.27405/d.cnki.gxbdu.2021.000162.
- [29] 陈道坤, 周海, 华红梅等.BP 神经网络和随机森林预测土壤有机质模型研究[J].安徽农学通报, 2023, 29(10):124-128.DOI:10.16377/j.cnki.issn1007-7731.2023.10.036.
- [30] Siewert, M. B.: High-resolution digital mapping of soil organic carbon in permafrost terrain using machine learning: a case study in a sub-Arctic peatland environment, Biogeosciences, 15,1663–1682, <https://doi.org/10.5194/bg-15-1663-2018>, 2018.

- [31] Mundada, Shyamal, Jain, Pooja, and Kumar, Nirmal. ‘Prediction of Soil Organic Carbon Using Machine Learning Techniques and Geospatial Data for Sustainable Agriculture’. 1 Jan. 2024 : 1 – 14.
- [32] 郑宇桐. 山西省土壤有机碳密度空间格局及其影响因素识别[D].西北农林科技大学, 2023.DOI:10.27409/d.cnki.gxbnu.2023.001392.
- [33] 高鹏利, 任大陆, 冯志强, 等.基于 Boruta 算法和 GA 优化混合地统计模型的土壤有机质空间分布预测[J/OL].物探与化探:1-12[2024-05-16].<http://kns.cnki.net/kcms/detail/11.1906.p.20240510.1116.002.html>.
- [34] 山西省自然资源厅.山西公布第三次国土调查数据 全省耕地 5800 余万亩[EB/OL]. (2022-01-28) [2024-05-28].[https://zrzyt.shanxi.gov.cn/xw/chnl399/202201/t20220128\\_4696572.shtml](https://zrzyt.shanxi.gov.cn/xw/chnl399/202201/t20220128_4696572.shtml)
- [35] 彭守璋. (2020). 中国 1km 分辨率逐月降水量数据集 (1901-2022) . 国家青藏高原数据中心. <https://doi.org/10.5281/zenodo.3185722>.
- [36] 彭守璋. (2019). 中国 1km 分辨率逐月平均气温数据集 (1901-2022) . 国家青藏高原数据中心. <https://doi.org/10.11888/Meteoro.tpdc.270961>. <https://cstr.cn/18406.11.Meteoro.tpdc.270961>.
- [37] 彭守璋. (2022). 中国 1km 逐月潜在蒸散发数据集 (1901-2022) . 国家青藏高原数据中心. <https://doi.org/10.11866/db.loess.2021.001>.
- [38] 叶治山. 基于 Google Earth Engine 和机器学习的土壤有机质含量空间分布预测和制图 [D].安徽农业大学, 2023.DOI:10.26919/d.cnki.gannu.2022.000203.
- [39] Pinzon, J.E., E.W. Pak, C.J. Tucker, U.S. Bhatt, G.V. Frost, and M.J. Macander. 2023. Global Vegetation Greenness (NDVI) from AVHRR GIMMS-3G+, 1981-2022. ORNL DAAC, Oak Ridge, Tennessee, USA. <https://doi.org/10.3334/ORNLDaac/2187>
- [40] Webster R, Oliver M. Geostatistics for environmental scientists[M]. New York: John Wiley & Sons, Ltd, 2001.
- [41] 杜睿山, 黄玉朋, 孟令东, 等.基于 BiLSTM-XGBoost 混合模型的储层岩性识别[J/OL]. 计算机系统应用:1-9[2024-05-17].<https://doi.org/10.15888/j.cnki.csa.009522>.
- [42] 周珺. 长期施肥对设施番茄土壤有机碳组分及固碳效应的影响[D].沈阳农业大学, 2023. DOI:10.27327/d.cnki.gshnu.2023.000378.
- [43] 白重九. 1980-2010 年北方旱地农田土壤有机碳变化特征及其主控因素研究[D].中国农业科学院, 2021.DOI:10.27630/d.cnki.gznky.2021.000716.
- [44] 杨雪茹. 松嫩草地土壤氮磷比对植物生长和种间相互作用的影响[D].东北师范大学, 2024.DOI:10.27011/d.cnki.gdbsu.2023.000703.
- [45] Allen, R.G., Pereira, L.S., Raes, D., & Smith, M. (1998). Crop evapotranspiration : guidelines for computing crop water requirements.

- [46] 王利平,文明,宋进喜,等.1961—2014 年中国干燥度指数的时空变化研究[J].自然资源学报,2016,31(09):1488-1498.
- [47] 张乾,杨姗姗. 1958—2019 年山西省降雨和气温时空变化特征分析[C]//中国水利学会.2023 中国水利学术大会论文集 (第三分册).黄河水利出版社,2023:9.DOI:10.26914/c.cnkihy.2023.088261.
- [48] 孙从建,郑振婧,李伟,等.1964-2017 年山西省潜在蒸发量时空变化及其影响因素分析[J].水土保持研究,2019,26(05):229-235.DOI:10.13869/j.cnki.rswc.2019.05.034.
- [49] 常大海,常仁凯.2000—2015 年山西省植被覆盖时空演变特征分析[J].水资源开发与管理,2023,9(04):38-44+37.DOI:10.16616/j.cnki.10-1326/TV.2023.04.08.