

基于因果推断和机器学习的农田玉米产量预测模型

王毅^{1,6}, 崔茜彤¹, 王晨¹, 熊宝伟¹, 邵国敏², 王琬莹¹,
曹培^{3*}, 韩文霆^{4,5*}

(1. 西安财经大学信息学院, 陕西西安 710100, 中国; 2. 西安理工大学西北旱区生态水利国家重点实验室, 陕西西安 710048, 中国; 3. 西北农林科技大学机械与电子工程学院, 陕西杨凌 712100, 中国; 4. 西北农林科技大学旱区农业节水研究院, 陕西杨凌 712100, 中国; 5. 国家作物高效用水工程实验室, 陕西杨凌 712100, 中国; 6. 智财协同可信计算陕西省高等学校重点实验室, 陕西西安 710100, 中国)

摘要: [目的/意义] 玉米作为中国主要粮食作物, 其产量的精准预测在维护国家粮食供给体系稳定方面具有重要意义。针对传统预测方法难以揭示变量间因果关系的问题, 提出了一种融合因果推断与机器学习的预测方法。[方法] 首先, 基于遥感指数、气象数据、土壤剖面湿度和阶段性作物观测数据等多源异构信息, 采用偏相关与瞬时条件独立算法系统分析玉米产量与各影响因子之间的因果关系。其次, 构建了移动平均-卷积神经网络-长短期记忆网络混合模型, 先通过移动平均模块平滑数据噪声, 进而利用卷积神经网络提取变量间的空间关联特征。最后, 通过长短期记忆网络捕捉产量随时间变化的动态特征。[结果和讨论] 10 cm 和 50 cm 深度的土壤湿度对产量具有显著的正向影响 ($P < 0.01$), 其中深层湿度具有更强的时滞效应; 反射率中的改进叶绿素吸收指数与归一化植被指数等植被指数则在玉米生长中期呈现出短期因果关系。与其他模型的对比实验结果表明, 所提出的模型在多个指标上表现突出, 其中决定系数 (R^2) 达到 0.955, 平均绝对误差和均方根误差分别为 1.201 kg/亩和 1.474 kg/亩 (1 hm²=15 亩)。[结论] 因果分析可为模型提供关键变量筛选依据, 显著提升产量预测性能。

关键词: 偏相关与瞬时条件独立算法; 移动平均; 卷积神经网络; 长短期记忆网络; 产量预测; 机器学习; 因果推断

中图分类号: S513.01; TP181; O212.1

文献标识码: A

文章编号: SA202506027

引用格式: 王毅, 崔茜彤, 王晨, 熊宝伟, 邵国敏, 王琬莹, 曹培, 韩文霆. 基于因果推断和机器学习的农田玉米产量预测模型[J]. 智慧农业(中英文), 2026, 8(2): 175-187. DOI: 10.12133/j.smartag.SA202506027

WANG Yi, CUI Xitong, WANG Chen, XIONG Baowei, SHAO Guomin, WANG Wanying, CAO Pei, HAN Wenting. Field Maize Yield Prediction Model Based on Causal Inference and Machine Learning in Agricultural Fields[J]. Smart Agriculture, 2026, 8(2): 175-187. DOI: 10.12133/j.smartag.SA202506027 (in Chinese with English abstract)

0 引言

粮食生产对国民经济至关重要, 玉米作为中国主要粮食作物之一, 对农业和经济发展具有显著影响。尽管中国玉米产量在过去 10 年间增长了 8.58%, 但仍存在较大的供需缺口。在此背景下, 探究玉米产量的变动规律及其影响因素, 在优化农业政策、提升产量、保障国家粮食安全方面具有关

键意义。然而, 作物产量的形成并非一蹴而就, 而是一个多步骤、多因素交织的复杂过程。它涉及植物的多个生理和生化反应, 受到诸如气候、土壤质量、田间管理等多种因素的影响^[1]。要了解产量的形成, 就需要明确作物生长各阶段的特性、这些阶段与产量之间的联系, 以及作物生长和产量对环境因素的响应。近年来, 机器学习进入快速发展时期, 与传统方法相比, 其在样本上的拟合稳定性也

收稿日期: 2025-06-17

基金项目: 国家社会科学基金 (23BGL252); 陕西省自然科学基金基础研究计划项目 (2022JQ-363); 陕西省重点产业创新链项目 (2024NC-ZD-CYL-05-01); 陕西省重点产业创新链项目 (2023-ZDLNY-58); 西安市科技计划项目 (25NJSYB00014)

作者简介: 王毅, 博士, 讲师, 研究方向为农情信息天空地一体化智能感知与精准作业技术。E-mail: wang_yi@xaufe.edu.cn

*通信作者: 1. 曹培, 博士, 副教授, 研究方向为农业水信息天空地一体化智能感知与精准灌溉技术及装备。E-mail: caopei2020@nwfufu.edu.cn; 2. 韩文霆, 博士, 研究员, 研究方向为农业水信息天空地一体化智能感知与精准灌溉技术及装备。E-mail: hanwt2000@126.com

copyright©2026 by the authors

(登录 www.smartag.net.cn 免费获取电子版全文)

更为优越。其中,长短期记忆网络(Long Short-Term Memory, LSTM)^[2]因具有较好地保存记忆时间序列数据的能力而被广泛应用于各行业的时序预测。王旭等^[3]通过门控机制提取了冬小麦生长过程中的时间序列特征,建立了基于LSTM的冬小麦估产模型。SHI等^[4]将LSTM与工业应用相结合,通过多传感器数据融合实现了高精度的振动预测。董莉霞等^[5]利用霍德里克-普雷斯科特滤波(Hodrick-Prescott, HP)算法对气象驱动产量与长期变化产量进行区分处理,并通过5种机器学习算法对气象产量进行预测,同时构建差分自回归移动平均模型(Autoregressive Integrated Moving Average Model, ARIMA)时间序列模型对趋势产量进行预测,最终通过LSTM和ARIMA模型的组合,实现了对旱地春小麦产量的高精度预测。YU等^[6]将卷积神经网络(Convolutional Neural Network, CNN)和LSTM网络融合,使系统能够在旋转过程中跟踪有缺陷的机器人并识别其真实类型,并预测其未来路径。王得道等^[7]构建了一种融合CNN与LSTM的预测模型,其中CNN用于提取气象条件对当前时刻板温的影响特征,LSTM用于预测,有效预测了中国铁路轨道系统(China Railway Track System, CRTS)Ⅱ型轨道板温度。WANG等^[8]提出了一种基于麻雀搜索算法(Sparrow Search Algorithm, SSA)优化的CNN-LSTM的混合模型,该模型通过引入CNN的空间表征提取优势,并融合LSTM在时间动态建模方面的能力,通过SSA进行全局优化,提高了多特征时间序列的预测精度。尽管上述模型在多特征时间序列建模中展现出较高的预测能力,但多数仅从历史数据中学习变量间的统计相关性,忽略了潜在的滞后因果关系,在农业遥感建模实践中,这一问题亦未得到有效解决。JOSHI等^[9]指出,传统统计模型虽然易于实现,但难以捕捉变量间复杂的非线性交互作用,依赖人工特征构建,且无法揭示驱动作物产量变化的因果机制。正如VAN CLEEMPUT等^[10]所强调,仅依赖遥感数据的传统方法难以揭示因果关系,且易受混杂变量与测量误差影响而产生偏差。这些研究表明,利用相关性建模方法虽取得一定进展,但仍难以揭示变量间的深层驱动机制。RUNGE^[11]指出,在高维、多变量、非线性时间序列中,传统方法易产生“伪相关”,误将非因果关联纳入建模依据,从而影响模型稳定性与可解释性。在该领域中,RUNGE等^[12]提出基于条件独立性的瞬时因果推断算法(Peter-

Clark and Momentary Conditional Independence, PC-MCI),通过结合用于条件选择的彼得-克拉克算法第一阶段(Peter-Clark 1, PC1)与瞬时条件独立性检验(Momentary Conditional Independence, MCI),系统性应对了自相关、非线性和高维等挑战,实现了大规模时间序列因果网络的高效重构。

基于上述问题,本研究首先引入PCMCI因果推断算法,以增加模型可解释性;其次在CNN-LSTM模型基础上引入移动平均(Moving Average, MA)模块,将MA预测结果作为额外的特征,使模型可以同时考虑原始特征和基于历史数据的趋势信息,在一定程度上优化了CNN-LSTM模型,提高了模型的预测精度;最后提出基于因果推断和机器学习的农田玉米产量预测模型。

1 材料与方法

1.1 实验地概况

本研究以华北平原北部的典型灌溉高产农业区——河北省定兴县固城镇农业气象野外科学实验基地为研究区域。该试验基地位于北纬39°08′、东经115°48′,海拔15.2 m,占地面积9.5 hm²。该区域位于典型的暖温带大陆性季风气候背景下,年均气温为12.2℃,年均降水量约为528 mm,主要集中在6—9月;年平均日照时数为2 264 h。近年来,该区域地下水位基本稳定在15 m左右。

试验场地平整,四周设有2 m宽的保护带,且周围均为相同土地利用类型的农田。试验场内采用常规施肥和喷灌方式灌溉,能够有效防御旱涝灾害和农业病虫害。样地总面积780 m² (30 m×26 m),共划分为16个小区,每个小区占地面积为30 m² (6 m×5 m),种植10行,每行18株;保护行设在样地周边,宽度2 m,样地边设有试验标牌。选用的玉米品种为廉玉1号,种植密度为60 003株/hm²,行距50 cm,株距33.3 cm。每个小区施底肥量2.5 kg,复合肥品牌为绿得利(N-P₂O₅-K₂O, 26-14-5,总养分≥45%),播种时肥料一次性施入。病虫害防治则采用了常规的农药防治措施。

在A、B、C、D共4个播期试验中,所有观测工作均在小区A1~A3、B1~B3、C1~C3、D1~D3中进行。而A4、B4、C4、D4共4个区域则专门用于产量要素的测定,不进行任何其他观测。这样的设计既保证了试验的科学性,又为后续的产量分析奠定了可靠的数据基础。分期播种试验场不同播期样地分布如图1所示。

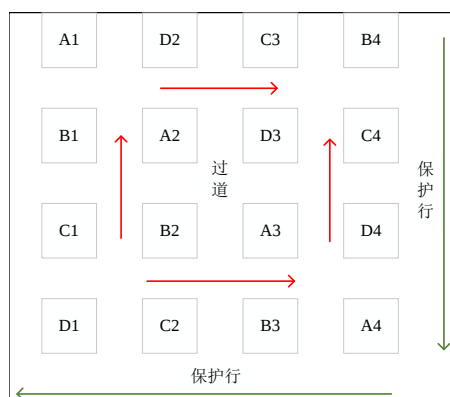


图1 分期播种试验场不同播期样地分布图

Fig. 1 Distribution map of different sowing periods' plots in the split-seeding experimental field

1.2 数据来源

农田玉米生长与当地气候、土壤温度及其他环境因子密切相关，如降水量、日照时数、温度等，会影响农田玉米的整个生长周期。因此本研究选用作物数据、气象数据、土壤数据和遥感数据4类影响因素（表1）。

1.2.1 作物及土壤数据

作物及土壤数据来源于科学数据银行（Science Data Bank）的“2018—2021年华北北部不同播期对玉米生长发育的影响数据集”（<https://www.scidb.cn/>）^[13]，包含株高、叶面积指数（Leaf Area Index, LAI）及10~50 cm分层土壤湿度（空间分辨

表1 玉米产量影响因素

Table 1 Factors influencing maize yield

数据类型	观测指标
作物数据	株高/cm、叶面积指数
气象数据	地面气压/hPa、平均气温/(°C)、降水量/mm、露点温度/(°C)、相对湿度/%、经向风速(U, m/s)、纬向风速(V, m/s)、太阳辐射净强度(net, J/(m²/d))、太阳辐射总强度(down, J/(m²/d))
土壤数据	土壤湿度(深度:10、20、30、40、50 cm)
遥感数据	增强植被指数(Enhanced Vegetation Index, EVI)、绿光归一化植被指数(Green Normalized Difference Vegetation Index, GNDVI)、归一化植被指数(Normalized Difference Vegetation Index, NDVI)、土壤调整植被指数(Soil-Adjusted Vegetation Index, SAVI)、转换叶绿素吸收比率指数(Transformed Chlorophyll Absorption in Reflectance Index, TCARI)、视觉大气阻力指数(Visible Atmospherically Resistant Index, VARI)、土壤饱和度(Simple Ratio, SR)、归一化差值红边(Normalized Difference Red Edge Index, NDRE)、归一化差分红边指数(Normalized Difference Red Edge Index, NDREI)、反射率中的改进叶绿素吸收指数(Modified Chlorophyll Absorption in Reflectance Index, MCARI)

率90 m)，数据基于河北省定兴县固城农业气象野外科学实验基地的田间观测获取。

1.2.2 遥感数据

本研究通过计算地面光谱仪采集的地物光谱反射率（来自前文数据集）获取植被指数。表2中包含了本研究所选取的植被指数，包括EVI、GNDVI、NDVI、SAVI、VARI、SR、TCARI、NDRE、MCA-RI，其计算如表2中公式（1）~公式（9）所示。

表2 植被指数及其计算公式

Table 2 Vegetation index and their calculation formulas

植被指数	公式	引用文献
EVI	$EVI = 2.5 \times (R_{NIR} - R_R) / (R_{NIR} + 6 \times R_R - 7.5 \times R_B + 1)$	(1) [14]
GNDVI	$GNDVI = (R_G - R_R) / (R_G + R_R)$	(2) [15]
NDVI	$NDVI = (R_{NIR} - R_R) / (R_{NIR} + R_R)$	(3) [16]
SAVI	$SAVI = (1 + L) \times (R_{NIR} - R_R) / (R_{NIR} + R_R + L)$	(4) [17]
TCARI	$TCARI = 3 \times [(R_{RE} - R_R) - 0.2 \times (R_{RE} - R_G) \times (R_{RE} / R_R)]$	(5) [18]
VARI	$VARI = (R_G - R_R) / (R_G + R_R - R_B)$	(6) [19]
SR	$SR = R_{NIR} / R_R$	(7) [20]
NDRE	$NDRE = (R_{NIR} - R_{RE}) / (R_{NIR} + R_{RE})$	(8) [21]
MCARI	$MCARI = [(R_{RE} - R_R) - 0.2 \times (R_{RE} - R_G)] \times (R_{RE} / R_R)$	(9) [22]

注： R_{NIR} 为近红外波段， R_R 为红光波段， R_B 为蓝光波段， R_G 为绿光波段， R_{RE} 为红边波段，参数 L 为土壤调节系数（通常取0.5）。

1.2.3 气象数据

气象因素是影响作物生长发育的核心因素之一，水分、光照与热量等关键生长条件均与气象过程密切相关。其中，温度通过调控光合作用与呼吸作用影响作物的物质积累效率，并在不同生育阶段调节其生长进程与产量形成潜力。光照是光合作用

的能量来源，可通过影响光合效率与同化物供给水平，进而作用于干物质的积累与分配。降水是作物重要的水分来源，能够维持植株水分状态并促进养分吸收与运输，为群体生长与产量形成提供保障。因此，温度、光照与降水等要素协同作用于作物生长过程与产量形成，是产量研究中不可或缺的基础

信息。

本研究数据集采用欧洲中期天气预报中心(European Centre for Medium-Range Weather Forecasts, ECMWF)的ERA5-Land日尺度数据集^[23],包括2018—2021年6—9月不同生育期的地面气压、平均气温、降水量、露点温度、相对湿度、经向风速、纬向风速、太阳净辐射、太阳总辐射等指标(空间分辨率 $0.1^{\circ} \times 0.1^{\circ}$)。

1.2.4 数据预处理

鉴于各类影响因子的数据规模庞大且分散存储于不同的数据平台,直接使用原始数据需要面临大量复杂且重复的预处理工作。为此,本研究对多源数据进行系统化整合处理:首先,针对玉米生长周期(6月初至9月底),选取2018—2021年6—9月的日尺度气象数据,并按七叶期、拔节期、抽雄期及乳熟期(具体的生育时期如表3所示),提取对应时段的土壤数据、作物数据及遥感数据;其次,对土壤湿度数据中存在的少量缺失值,采用线性插值法填补以保持时间序列连续性;最后,通过双线性插值将土壤数据与气象数据统一至1 km网格尺度,确保空间匹配误差 $\leq 5\%$ 。通过上述处理,整合后的数据集涵盖12个关键影响因子(经PCMCI算法筛选),为后续建模提供了标准化、高可靠性的输入基础。

表3 不同研究区域玉米的关键生育期

Table 3 Key growth stages of maize in different research regions

名称	七叶期	拔节期	抽雄期	乳熟期
A区	6月25日	7月3日	7月28日	8月22日
B区	7月4日	7月10日	8月4日	8月27日
C区	7月13日	7月21日	8月13日	9月5日
D区	7月23日	7月30日	8月22日	9月18日

1.3 研究方法

1.3.1 PCMCI算法与参数设置

PCMCI算法是一种时间序列因果推断的算法,它是PC算法的扩展,用于在大型非线性时间序列数据集中检测和量化因果关联。其能够处理时间序列,捕捉变量之间的时间滞后效应。算法的关键在于独立性检验阶段,它通过统计检验判断变量之间是否独立,从而控制假阳性错误,更准确地估计因果效应的大小。

PCMCI算法分为两个主要阶段:PC1条件选择阶段和MCI因果发现阶段。首先,使用PC算法生成骨架,确定每个变量的初步父节点集。其次,通过MCI对每对变量进行进一步判断,以确定因果关

系的方向和强度。

1) 条件选择(PC1算法)。PC1算法是一种快速变体,用于估计每个变量 X_i^t 的父集 $P(X_i^t)$ 。具体步骤为:对于每个变量,初始化1个包含所有可能滞后变量的初步父集;从 $P=0$ 开始,逐步增加 P 值,对每个变量,测试其初步父集中的变量是否在给定其他变量的条件下与之独立;若某个变量在条件独立性测试中被判定为独立,则将其从初步父集中移除;当父集不再改变或迭代次数达到预设上限时,算法收敛,得到最终的父集估计。

2) MCI因果发现。在获得变量的父集后,MCI测试用于评估变量对之间的条件独立性,从而确定因果关系。其步骤为:对于所有变量对 (X_i^t, X_j^t) 和时间延迟 τ ,构建条件集,该条件集包含 X_i^t 的父集,以及 $X_j^{t-\tau}$ 的父集;进行条件独立性测试,若拒绝原假设,则认为存在因果关系;可选地调整 P 值以控制错误发现率。

3) 条件独立性测试(Conditional Independence Test)。是PCMCI算法的核心步骤之一,其目的是从高维、非线性且可能存在混杂因素的时间序列数据中,识别变量间的直接因果关系,并排除虚假关联。对于变量 X 、 Y 和条件集 Z ,若 $P(X, Y|Z) = P(X|Z)P(Y|Z)$,则称 X 、 Y 在给定 Z 的条件下独立。在因果推断中,条件独立性测试用于判断两个变量之间的关联是否由直接因果作用引起,而非通过其他变量间接关联(即阻断后门路径)。在PCMCI中,常用的条件独立性测试方法包括:偏相关(Partial Correlation, Par Corr)、高斯过程距离相关(Gaussian Process Distance Correlation, GPDC)和条件互信息(Conditional Mutual Information, CMI)^[12]。

偏相关适合分析线性关系,计算简单,但仅限于线性依赖。高斯过程距离相关适合分析非线性关系,计算复杂但鲁棒性强。考虑到本研究数据的非线性特征,以及数据量的有限性,选用GPDC用于本研究的条件独立性测试。

PCMCI算法的核心参数包括最大时间延迟($\tau_{\max}$)、显著性水平(Alpha)及条件集大小(P),其中, $\tau_{\max}$ 与Alpha对本研究的因果关系识别尤为关键。

显著性水平作为假设检验中拒绝零假设的临界阈值,决定了因果关系识别的严格程度:取值越小(如0.01),对“不存在因果关系”的零假设拒绝越谨慎,可有效降低假阳性风险(错误识别不存在的

因果关系)；取值较大(如0.05)则更易捕捉潜在关联，但假阳性概率升高。结合本研究对“各变量与玉米产量(Yield)因果关系准确性”的核心诉求，实验将显著性阈值(pc_alpha)设定为0.01，以严格控制假阳性，确保识别的因果关系更可靠^[24]。

最大时间延迟(tau_max)用于定义因果关系的时间滞后范围，其取值需结合研究对象的时间尺度与科学问题确定。为系统探究各变量对产量的即时影响(当前时刻)与滞后效应(过去不同时期)，本实验设置tau_max分别为0、1、2、3(对应不同生育期的0~3个时间滞后)，通过对比不同滞后条件下的因果关系结果，明确各变量影响产量的时间尺度特征(如是否为即时效应、短期滞后效应或中长期累积效应)。实验结果通过热力图可视化各变量在不同滞后条件下与产量的关联显著性(P值)，直观呈现不同时间尺度下的因果关系强度与显著性差异。

1.3.2 CNN-LSTM模型构建

通过一维卷积操作获取时间序列数据中的局部特征，捕捉短期依赖关系，LSTM作为循环神经网络的一种变体，主要面向时间序列数据建模任务。其通过记忆单元配合门控机制(包括输入门、遗忘门和输出门)实现对长期依赖关系的有效建模。

CNN依托卷积运算机制，对时间序列中的局部特征进行自动提取，并通过降维减少数据的复杂度。提取的特征随后被传递给LSTM层，进一步捕捉时间序列中的长期依赖关系^[8]。具体模型结构如图2所示。

1.3.3 MA模型

MA模型属于时间序列分析领域的经典方法，其核心在于刻画当前观测值与历史随机扰动项之间的关联结构。该模型建立在白噪声序列假设基础之上，其中白噪声被视为一种特殊形式的时间序列，其各时刻取值相互独立、服从一致分布特征，并具有固定的均值与方差。

q 阶移动平均模型通常简记为MA(q)，如公式(10)所示。

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \quad (10)$$

式中： y_t 为在时间点 t 的标签值； μ 为当前时间序列标签的均值； $\theta_1, \dots, \theta_q$ 为MA(q)模型的移动平均系数， q 为模型阶数； ε_t 为在时间点 t 时，不可预料、不可估计的偶然事件的影响； ε_{t-1} 为在时间点 $t-1$ 时不可预料、不可估计的偶然事件的影响；

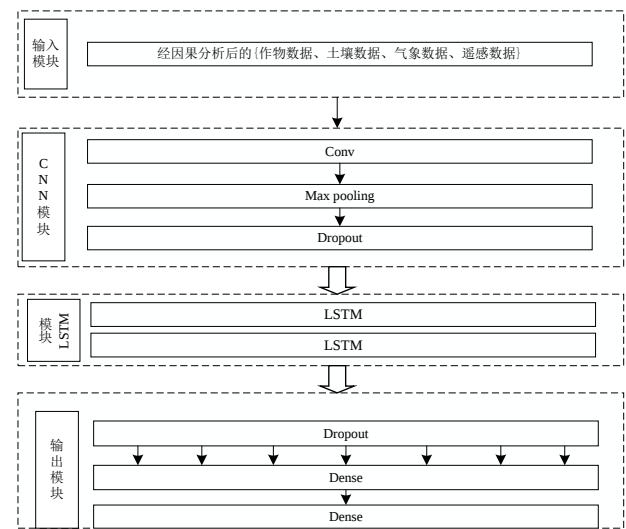


图2 CNN-LSTM模型结构图

Fig. 2 Structure diagram of CNN-LSTM model

$\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-q}$ 为白噪声^[25]。

1.3.4 MA-CNN-LSTM模型构建

本研究在CNN-LSTM组合结构中引入移动平均模块，构建MA-CNN-LSTM预测框架，以平滑多源时间序列数据的短期波动并强化趋势信息表征。研究中使用的植被指数与气象要素等序列数据包含观测噪声与随机扰动，使输入信号的短期波动成分增强，从而影响深度模型对关键规律的学习。移动平均采用滑动窗口对序列进行均值化处理，使序列更加平稳并突出局部趋势。对于时间点 t ，其移动平均值定义为该时间点及其前 k 个时间点观测值的均值，由此得到反映局部趋势特征的平滑序列，为后续特征学习提供更稳定的输入基础。

在模型流程上，原始多源时间序列首先经过移动平均模块获得平滑序列；其次将平滑序列输入CNN，通过卷积核在时间维上滑动提取局部模式特征，实现对阶段性变化与局部结构信息的有效编码，并表征多变量之间的局部关联关系；最后将CNN输出的特征序列输入LSTM，利用门控机制建模时间依赖关系，学习跨时间尺度的长期依赖与动态演化规律，并通过输出层生成最终预测值。在此基础上，移动平均用于平滑与稳定输入，CNN负责提取局部模式与关联特征，LSTM负责刻画长期依赖与时序动态，三者协同构成MA-CNN-LSTM预测框架，从而在一定程度上减弱短期波动对训练与预测的影响，并提升预测结果的稳定性。模型预测的整体流程如图3所示。

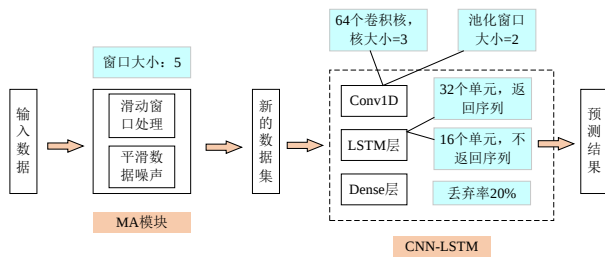


图3 MA-CNN-LSTM模型预测流程图

Fig. 3 Flowchart of the MA-CNN-LSTM model prediction process

1.3.5 MA-CNN-LSTM模型预测流程图

本研究围绕农业产量预测建模问题,构建了一套融合数据集成、因果推断与深度学习优化的综合技术路径。首先,数据收集阶段获取了作物、气象、土壤和遥感数据。其次,通过数据预处理将数据清洗并转化为适合模型输入的格式。再次,利用因果推断分析了影响玉米产量的关键因素,提供了模型的输入变量。在机器学习阶段,使用了初始的CNN-LSTM模型来处理数据的空间和时间依赖性,并通过引入MA-CNN-LSTM改进模型,利用移动平均平滑数据以提高预测精度和稳定性。最后,模型通过平均绝对误差(Mean Absolute Error, MAE)、均方根误差(Root Mean Square Error, RMSE)和决定系数(Coefficient of Determination, R^2)等评估指标进行性能评估,并通过模型解释分析,形成了从因果推断到预测验证的完整闭环。玉米产量预测的完整流程如图4所示。

1.3.6 模型评价指标

为评估预测模型的泛化性能,本研究采用

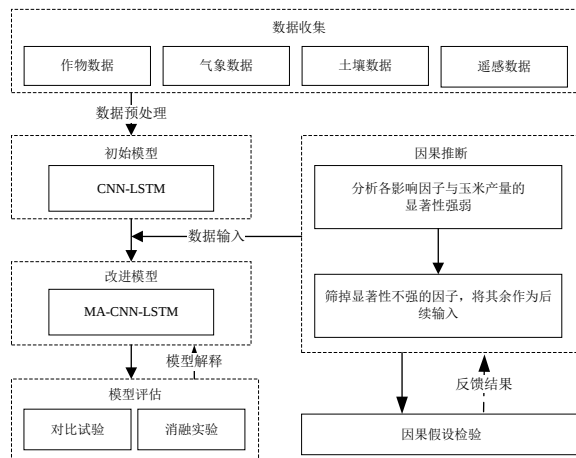


图4 基于因果推断和机器学习的农田玉米产量预测模型研究技术路线图

Fig. 4 Technical roadmap for research on farmland maize yield prediction model based on causal inference and machine learning

MAE、RMSE和 R^2 作为评价指标。MAE用于定量评估预测结果与实际观测值的偏离程度,该指标通过对所有样本点绝对误差取平均得到,其中绝对误差定义为预测值与真实值之间的非负差值。计算如公式(11)所示。

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (11)$$

式中: n 为数据点的总数; y_i 为第*i*个数据点的实际观测值; $\hat{y}_i$ 为第*i*个数据点的预测值; $|\cdot|$ 为绝对值。

RMSE属于常见的误差评估指标,常用于度量模型预测值与实际观测值之间的偏差大小。它是所有数据点的平方误差的平均值的平方根,可以提供误差的大小,并且与原始数据在同一量纲上。计算如公式(12)所示。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (12)$$

式中: $(y_i - \hat{y}_i)^2$ 为预测值与真实值之间的差的平方。

R^2 是指决定系数,是一个统计指标,用于表示因变量的方差中有多少比例可以从自变量预测得到,计算如公式(13)所示。

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (13)$$

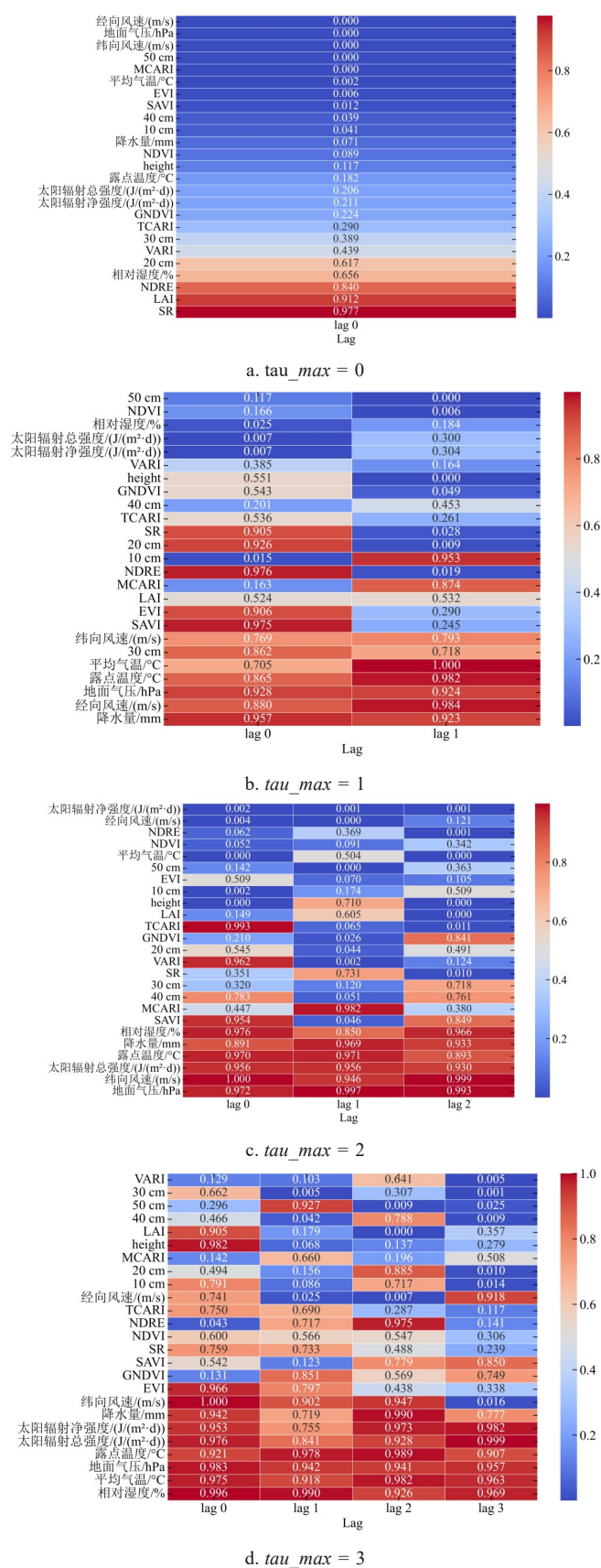
当 R^2 取值逐渐接近1时,表明模型的预测精度相应增强,即模型能够更好地解释数据中的变异性。如果 R^2 的值接近0,则表示模型的预测效果很差,几乎没有解释数据中的变异性。

2 结果与分析

2.1 PCMC1结果分析

玉米产量与各影响因子之间的显著性如图5所示。图例表示显著性水平:颜色越接近蓝色, P 值的数值越小,意味着该特征变量与目标变量之间的统计显著性越强,其相关关系也更为可靠;颜色越接近红色, P 值越大,显著性越弱。

图5a~图5d展示了不同时间滞后阶数,通过逐步增加时间滞后($0 \leq \tau_{max} \leq 3$)分析变量对作物产量的因果影响。若仅考虑即时效应($\tau_{max} = 0$),MCARI和50 cm土壤湿度等变量显著影响产量,反映植被光合效率和深层水分对产量的直接作用;而SR等变量因噪声或间接作用未呈现显著关



注: $P < 0.05$ 表示关系显著, $P < 0.01$ 表示高度显著, Lag 表示变量在不同时间滞后下的因果关系显著性差异。

图5 不同气象、作物、土壤、遥感因子与产量之间的 P 值

Fig. 5 P -value between different meteorological, crop, soil, remote sensing factors and yield

联。引入滞后效应后，不同变量的影响随时间变化：height 和 NDVI 在滞后 1 期 ($\tau_{\max}=1$) 显著，表明株高和植被指数存在延迟效应；10 cm 土壤湿度在滞后 2 期 ($\tau_{\max}=2$) 才显著，说明浅层水分的影响需要时间累积；而滞后 3 期 ($\tau_{\max}=3$) 时，深层土壤湿度和 VARI 等变量的长期效应更明显，但气象因素（如降水、气温）因间接作用或效应被覆盖，始终未呈现强因果关系。整体表明，植被指数和土壤湿度对产量具有直接且持续的因果影响，而气象因素的作用可能通过中介变量间接实现。

根据上述分析，可以得出不同时间滞后下特征变量与目标变量的显著性，如表 4 所示。

表 4 玉米产量预测模型输入特征筛选

Table 4 Input feature screening for maize yield prediction models

数据类型	时间滞后			
	Lag 0	Lag 0—Lag 1	Lag 0—Lag 2	Lag 0—Lag 3
地面气压/hPa	▲			
平均气温/°C	▲		▲	
降水量/mm				
露点温度/°C				
相对湿度/%		●		
经向风速/(m/s)	▲		▲	▲
纬向风速/(m/s)	▲			●
太阳辐射净强度/(W/m ²)		▲	▲	
太阳辐射总强度/(W/m ²)		▲		
土壤湿度 10 cm	●	●	▲	●
土壤湿度 20 cm		▲	●	▲
土壤湿度 30 cm				▲
土壤湿度 40 cm	●			▲
土壤湿度 50 cm	▲	▲	▲	▲
株高/cm		▲	▲	
叶面积指数/cm ²				▲
EVI	▲			
GNDVI		●	●	
NDVI		▲		
SAVI	●		●	
TCARI			●	
VARI			▲	
SR		●		
NDRE		●		●
MCARI	▲			

注：▲代表显著性 $P<0.01$ ；●代表显著性 $P<0.05$ 。

从表 4 中可以看出，降水量和露点温度在所有滞后阶数 (Lag0~Lag3) 中均未表现出与玉米产量的显著因果关系。这表明其可能对产量的直接影响

较弱，或者其作用需要通过其他中介变量间接体现。为了确保后续预测模型的准确性和严谨性，本研究采用更严格的筛选标准，仅保留在任意滞后阶数下 P 值 <0.01 的显著影响因子（表中标记 ▲ 的变量），而剔除 P 值 >0.01 的变量（包括部分仅达到 $P<0.05$ 但未达到 $P<0.01$ 的指标，如相对湿度、GNDVI 等）。这一筛选过程不仅去除了噪声干扰，并且突出了植被生理参数和深层土壤水分对产量的主导作用，为后续预测模型提供了更可靠的特征集。

2.2 预测模型对比分析

为明确 MA 窗口大小对模型性能的具体影响，在原有消融实验基础上，新增不同窗口尺寸对比实验，设置窗口大小为 1（无平滑，即原始数据）、5、7，其余模型参数（如 CNN 卷积核、LSTM 单元数、Dropout 率）保持一致。实验以未加入 MA 模块的 CNN-LSTM 模型为基准，通过对比不同窗口下的预测精度 (R^2 、MAE、RMSE)，筛选最优窗口配置。具体实验结果如表 5 所示。

表 5 MA 窗口大小的梯度实验设计

Table 5 Gradient experimental design for MA window size

模型配置	MAE/(kg/亩)	RMSE/(kg/亩)	R^2	核心变化分析
无 MA (窗口=1)	1.798	2.105	0.908	原始数据含较多噪声，预测曲线波动明显
MA 窗口=5	1.564	1.924	0.923	平衡噪声过滤与特征保留，性能最优
MA 窗口=7	1.627	1.983	0.915	过度平滑导致关键时序特征（如抽雄期突增）被弱化

注：1 hm²=15 亩(下表同)。

为了评估 MA-CNN-LSTM 模型的泛化能力和性能表现，本研究设置了对比实验，以验证其相较于其他产量预测模型的优越性。

首先，将 MA-CNN-LSTM 模型分别与支持向量回归 (Support Vector Regression, SVR)、最小绝对收缩和选择算子 (Least Absolute Shrinkage and Selection Operator, LASSO) 回归、CNN，以及 LSTM 这 4 种预测模型进行对比分析。通过对比实验，验证 MA-CNN-LSTM 模型相较于其他预测模型在产量预测中的显著优势。

图 6a 展示了 MA-CNN-LSTM 模型的预测结果，将其与图 6b~图 6e 进行对比实验分析。

图 6b 中，SVR 模型预测曲线波动较大，尤其在测试样本数量为 5~10 和 25~30 时出现明显偏离，反映了 SVR 对非线性时序特征捕捉能力的不足。图 6c 中的 LASSO 回归预测结果整体较差，可

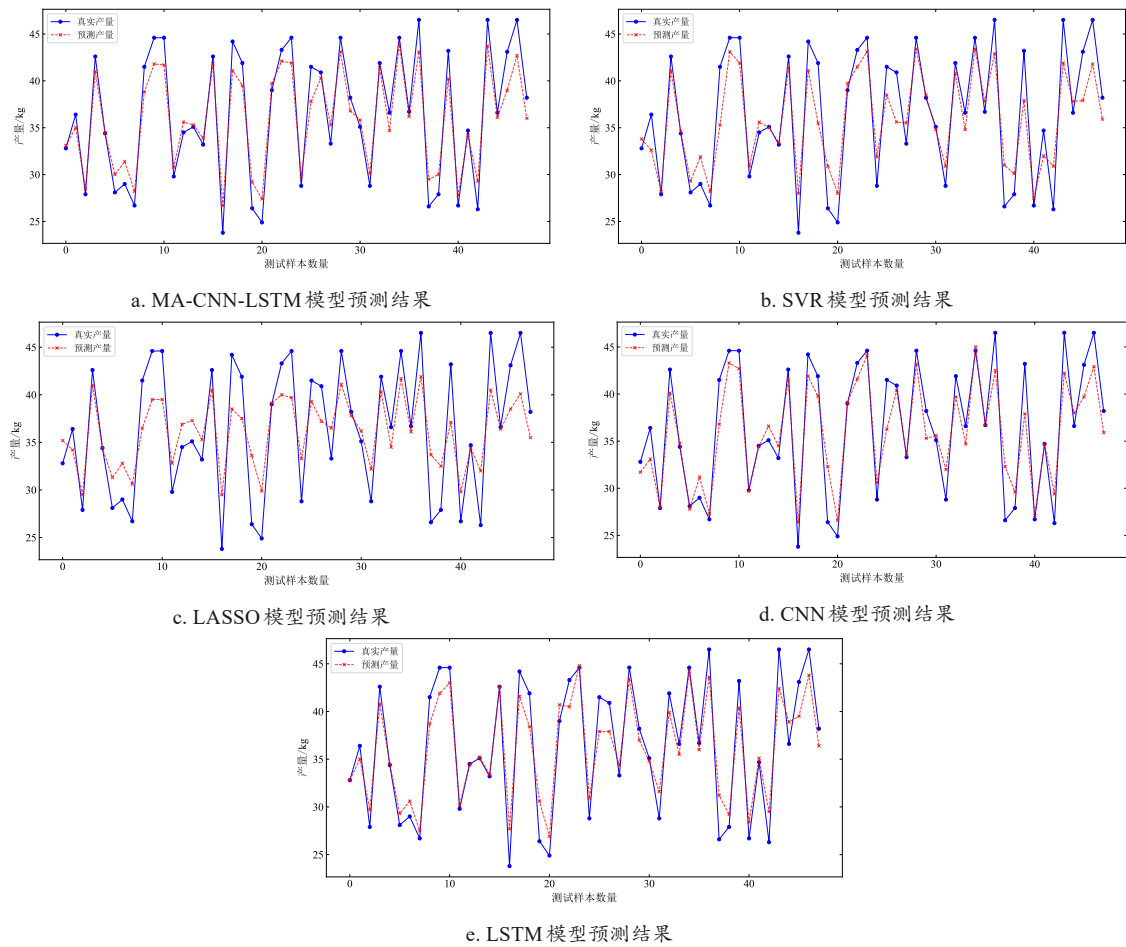


图6 不同模型预测玉米产量结果
Fig. 6 Maize yield prediction results of different models

能是由于目标变量与特征之间呈非线性关联，导致LASSO回归模型无法捕捉到目标，因此预测精度偏低。图6d中，CNN预测结果虽较SVR稳定，但对玉米生育期动态变化的响应滞后，例如在抽雄期（样本25附近）未能准确捕捉产量骤升趋势。图6e中，LSTM模型虽能部分捕捉长期依赖关系，但在极端气象事件（如样本5附近的干旱期）预测误差显著增大，表明其对瞬时噪声的敏感性。相比之下，MA-CNN-LSTM模型通过融合移动平均模块有效抑制了气象数据的随机波动（如径向风速的瞬时跳变），同时在关键生育期的预测与实际值几乎重合，验证了其时空特征提取与噪声抑制的双重优势。表6为对比试验模型评估。

从表6中的模型评估结果来看，MA-CNN-LSTM模型在预测玉米产量上明显优于其他模型。从测试集的指标来看，MA-CNN-LSTM模型的 R^2 达到了0.955，表明其能够解释约95.5%的数据变异性，远高于其他单一模型（如LASSO的0.679）。

表6 不同玉米产量预测模型评估结果
Table 6 Evaluation results of different maize yield prediction models

Model	MAE/(kg/亩)	RMSE/(kg/亩)	R^2
SVR	2.363	2.943	0.821
CNN	2.007	2.564	0.864
LASSO	3.473	3.947	0.679
LSTM	1.833	2.217	0.898
MA-CNN-LSTM	1.201	1.474	0.955

同时，MA-CNN-LSTM的MAE和RMSE分别为1.201 kg/亩和1.474 kg/亩（1 hm²=15 亩），均为所有模型中最低，这进一步证明了其在预测精度和误差控制方面的优越性。相比之下，单一的LASSO模型在预测精度和误差控制方面表现欠佳，而CNN、LSTM和SVR虽然在一定程度上优于LASSO，但仍然略逊于MA-CNN-LSTM。实验结果表明，MA-CNN-LSTM模型在拟合效果和预测精度方面显著优于其他对比模型，其各项评估指标均表现突出，同时展现出较强的稳定性。

2.3 消融实验

为探究CNN-LSTM模型所使用PCMCI因果推断的因子分析、MA模型的优化方式对模型带来的性能提升,设置消融实验。具体实验结果如图7和表7所示。

如图7a所示,CNN-LSTM预测结果中,预测曲线在样本数量为5~10和25~30出现显著波动,

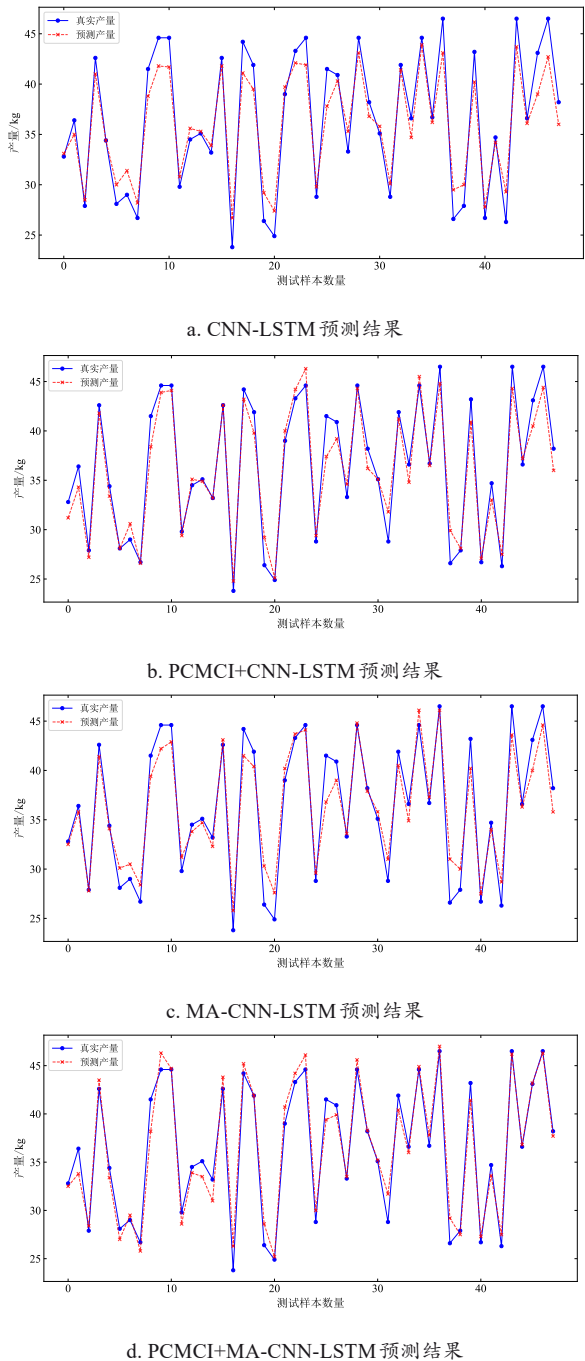


图7 不同玉米产量预测模型消融实验预测结果

Fig. 7 Ablation experiment prediction results of different maize yield prediction models

这表明原始CNN-LSTM模型对气象数据瞬时噪声和传感器随机误差较为敏感。相比之下,引入MA模块的MA-CNN-LSTM模型(图7c)在相同区间的预测曲线平滑度显著提升,这说明MA模块有效滤除了高频噪声,使模型更聚焦于生育期的趋势性特征。

对比图7c和图7d可发现,未引入PCMCI的MA-CNN-LSTM模型(图7c)在样本数量为15~20时出现系统性低估,这是由于传统模型输入中包含冗余变量导致模型对关键生育期水分胁迫信号的捕捉能力下降。此外,图7d的PCMCI+MA-CNN-LSTM模型通过因果网络筛选出核心变量,其预测曲线与真实产量一致性较高,证实PCMCI算法在提升产量预测准确性方面的积极作用。

由表7可知,融合PCMCI算法与MA-CNN-LSTM模型的 R^2 达到了0.955,在所有模型中最高,而CNN-LSTM模型的 R^2 值为0.908,相对较低。与CNN-LSTM模型相比,MA-CNN-LSTM模型的 R^2 值提升了1.6个百分点,说明CNN-LSTM虽然在融合空间和时间特征方面表现出色,但MA-CNN-LSTM通过引入移动平均机制进一步优化了模型性能,使其在泛化能力和预测精度上达到更高水平,而PCMCI+MA-CNN-LSTM模型在此基础上又提高了3个百分点,这说明PCMCI算法能够通过识别和强化关键的因果关系,进一步优化模型对数据特征的学习和提取,从而提高了预测的准确性。此外,通过比较MAE和RMSE指标,PCMCI+MA-CNN-LSTM模型的误差最低,进一步证实了该模型在提高玉米产量预测精度方面的显著优势。综上所述,本研究所提出的方法在提升玉米产量预测的准确性方面具有明显优势。

表7 玉米产量预测模型构建消融实验评估结果

Table 7 Ablation experiment evaluation results of maize yield prediction models

Model	Factor		MAE/ (kg/亩)	RMSE/ (kg/亩)	R^2
	PCMCI	MA			
CNN-LSTM	×	×	1.798	2.105	0.908
	√	×	1.359	1.691	0.941
	×	√	1.564	1.924	0.923
	√	√	1.201	1.474	0.955

注: ×代表不加入该因素; √表示加入该因素。

3 讨论与结论

3.1 讨论

作物产量的形成受气候、土壤、田间管理等多

因素共同作用，变量间的非线性与滞后关系在预测中具有重要意义。本研究构建的PCMCi+MA-CNN-LSTM模型在特征筛选与时序建模上进行了结合，在多源非平稳数据分析中能够同时考虑长期趋势与短期波动。在特征选择环节，PCMCi因果推断算法识别出MCARI、NDVI和深层土壤湿度等在多个滞后阶数下对玉米产量具有较强的统计显著性，这些因子在本研究数据中与作物光合效率、氮素营养和水分供应等生理机制相符。因果筛选在一定程度上减少了冗余特征，并增强了输入变量与产量变化之间的相关性，有助于提升模型的解释性与稳健性。

在时序建模方面，引入MA机制对遥感与气象时间序列中的高频波动进行平滑处理，使CNN-LSTM能够更好地提取季节性趋势与长期变化信号。在极端气象波动或观测噪声较大的情境下，该机制在一定程度上表现出稳定性与抗干扰能力。实验结果显示，集成因果变量筛选与MA模块的CNN-LSTM模型在本研究条件下的多项评估指标上取得了较优表现，并在样本扰动及阶段性变化的情境下表现出一定的适应性，说明该方法在多源非平稳农业数据分析中具有一定的应用潜力。

与既有研究相比，本研究在方法设计上实现了有益拓展。冯海宽等^[26]利用无人机高光谱数据结合逐步回归构建冬小麦氮含量模型。王超等^[27]基于PCA与PLSR建立综合长势指数监测模型。罗琦等^[28]融合气象、空间及灾害特征变量，利用GBDT、SVM、BRBP和MLR预测黄土高原苹果相对气象产量。赵泽阳等^[29]利用无人机多时相多特征数据结合PLSR、SVR、RF和BP等模型开展冬小麦产量预测，结果表明多生育时期融合的SVR模型精度较高。王来刚等^[30]基于多源时空数据与随机森林构建冬小麦产量预测模型。这些研究在光谱特征提取、指数构建、多源数据融合和机器学习建模等方面提供了重要参考。本研究在此基础上引入因果推断方法，结合深度学习的时序建模能力与趋势平滑机制，为多源数据驱动的作物产量预测提供了一种可行的思路与实现路径。

总体来看，该模型在因果变量筛选、时序特征提取及趋势平滑等方面形成了较为系统的解决方案，在本研究条件下表现出预测性能和可解释性的提升。但需要注意的是，本研究所用数据来源于特定区域和时间范围，模型的适用性和稳定性在更大空间尺度及跨作物类型的情境下仍有待进一步验证。未来可结合迁移学习与实时遥感、气象数据

流，推动该方法在多区域、动态化作物产量预测与预警中的应用，并探索轻量化与高效部署策略，以更好地服务于农业生产管理。

3.2 结论

本研究针对玉米产量预测中高维、多源及非线性特征建模的难题，提出了融合因果变量筛选与深度学习优化的预测框架——PCMCi+MA-CNN-LSTM。通过PCMCi因果推断算法，识别出MCARI、NDVI和深层土壤湿度等在多个滞后阶数下均显著影响产量的关键因子，提升了特征集的科学性与模型的解释力；在时序建模中引入滑动平均机制，有助于平滑非平稳序列的高频波动，使CNN-LSTM能够更精准地捕捉趋势信号与季节性变化。基于测试数据集的评估结果显示，该模型的 R^2 达到0.955，其MAE和RMSE分别为1.201 kg/亩和1.474 kg/亩，对比其他模型在预测精度与稳定性方面均有提升。消融实验验证了因果筛选与MA模块在性能优化中的积极作用。该研究为农业遥感多源数据驱动的作物产量预测提供了一条可行路径，在因果性挖掘与深度学习融合方面展现出良好的适用性与推广潜力。未来可在多作物、多区域及多时相数据上进一步验证模型的泛化能力，并结合模型压缩与高效部署策略，推动其在农业智能管理与产量动态预警中的实际应用。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献：

- [1] XIAO G L, HUANG J X, ZHUO W, et al. Progress and perspectives of crop yield forecasting with remote sensing: A review[J]. IEEE Geoscience and Remote Sensing Magazine, 2025, 13(3): 338-368.
- [2] 吴超, 周紫静, 黄锦铎, 等. 基于长短时记忆的农作物生长环境数据预测[J]. 深圳大学学报(理工版), 2024, 41(5): 563-573.
WU C, ZHOU Z J, HUANG J H, et al. Prediction of crop growth environment data based on Long Short-Term Memory[J]. Journal of Shenzhen University (Science & Engineering), 2024, 41(5): 563-573.
- [3] 王旭, 刘波, 陈正超, 等. 基于多源数据和LSTM模型的县域冬小麦估产[J]. 农业现代化研究, 2023, 44(6): 1117-1126.
WANG X, LIU B, CHEN Z C, et al. Winter wheat yield estimation at county-scale based on the multi-source data and LSTM model[J]. Research of Agricultural Modernization, 2023, 44(6): 1117-1126.
- [4] SHI Z L, DUAN J J, LI F Q. Machine tool operating vibration prediction based on multi-sensor fusion and LSTM neural network[J]. Electronics Letters, 2024, 60(22): e70100.
- [5] 董莉霞, 张博, 李广, 等. 基于机器学习算法和ARIMA模型的旱地春小麦产量预测[J]. 麦类作物学报, 2024, 44(12): 1551-1559.
DONG L X, ZHANG B, LI G, et al. Prediction of dryland spring

- wheat yield based on machine learning algorithms and ARIMA model[J]. *Journal of Triticeae Crops*, 2024, 44(12): 1551-1559.
- [6] YU Y H, AN X S, LIN J H, et al. A vision system based on CNN-LSTM for robotic citrus sorting[J]. *Information Processing In Agriculture*, 2024, 11(1): 14-25.
- [7] 王得道, 王森荣, 林超, 等. 基于CNN-LSTM融合神经网络的CRTS II型轨道板温度预测方法[J]. *铁道学报*, 2023, 45(2): 108-115.
WANG D D, WANG S R, LIN C, et al. Temperature prediction method for CRTS II type track slab based on CNN-LSTM integrated neural network[J]. *Journal of the China Railway Society*, 2023, 45(2): 108-115.
- [8] WANG J, XU D, YANG W S, et al. Advanced thermal prediction for green roofs: CNN-LSTM model with SSA optimization[J]. *Energy and Buildings*, 2024, 322: 114745.
- [9] JOSHI A, PRADHAN B, GITE S, et al. Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: A systematic review[J]. *Remote Sensing*, 2023, 15(8): 2014.
- [10] VAN CLEEMPUT E, ADLER P B, SUDING K N, et al. Scaling-up ecological understanding with remote sensing and causal inference[J]. *Trends in Ecology & Evolution*, 2025, 40(2): 122-135.
- [11] RUNGE J. Causal network reconstruction from time series: From theoretical assumptions to practical estimation[J]. *Chaos*, 2018, 28(7): 075310.
- [12] RUNGE J, NOWACK P, KRETSCHMER M, et al. Detecting and quantifying causal associations in large nonlinear time series datasets[J]. *Science Advances*, 2019, 5(11): eaau4996.
- [13] 耿金剑, 周广胜, 宋艳玲, 等. 2018—2021年华北北部不同播期对玉米生长发育的影响数据集[J]. *中国科学数据*, 2023, 8(4): 179-188.
GENG J J, ZHOU G S, SONG Y L, et al. A dataset of effects of different sowing dates on maize growth and development in the northern part of North China from 2018 to 2021[J]. *China Scientific Data*, 2023, 8(4): 179-188.
- [14] SHI H, LI L H, EAMUS D, et al. Assessing the ability of MODIS EVI to estimate terrestrial ecosystem gross primary production of multiple land cover types[J]. *Ecological Indicators*, 2017, 72: 153-164.
- [15] MAIMAITIJIANG M, SAGAN V, SIDIKE P, et al. Soybean yield prediction from UAV using multimodal data fusion and deep learning[J]. *Remote Sensing of Environment*, 2020, 237: 111599.
- [16] TESTA S, SOUDANI K, BOSCHETTI L, et al. MODIS-derived EVI, NDVI and WDRVI time series to estimate phenological metrics in French deciduous forests[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2018, 64: 132-144.
- [17] HERRERO-HUERTA M, RODRIGUEZ-GONZALVEZ P, RAINEY K M. Yield prediction by machine learning from UAS-based multi-sensor data fusion in soybean[J]. *Plant Methods*, 2020, 16(1): 78.
- [18] SHARIFI A. Remotely sensed vegetation indices for crop nutrition mapping[J]. *Journal of the Science of Food and Agriculture*, 2020, 100(14): 5191-5196.
- [19] ZHOU X, ZHENG H B, XU X Q, et al. Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017, 130: 246-255.
- [20] 张慧, 李平衡, 周国模, 等. 植被指数的地形效应研究进展[J]. *应用生态学报*, 2018, 29(2): 669-677.
ZHANG H, LI P H, ZHOU G M, et al. Advances in the studies on topographic effects of vegetation indices[J]. *Chinese Journal of Applied Ecology*, 2018, 29(2): 669-677.
- [21] 束美艳, 顾晓鹤, 孙林, 等. 基于新型植被指数的冬小麦 LAI 高光谱反演[J]. *中国农业科学*, 2018, 51(18): 3486-3496.
SHU M Y, GU X H, SUN L, et al. High spectral inversion of winter wheat LAI based on new vegetation index[J]. *Scientia Agricultura Sinica*, 2018, 51(18): 3486-3496.
- [22] QIAO L, TANG W J, GAO D H, et al. UAV-based chlorophyll content estimation by evaluating vegetation index responses under different crop coverages[J]. *Computers and Electronics in Agriculture*, 2022, 196: 106775.
- [23] MUÑOZ SABATER J. ERA5-Land monthly averaged data from 1981 to present[DS/OL]. [2025-06-12]. <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land-monthly-means?tab=overview>.
- [24] ZHAO C L, ZHU W Q, CHEN L Y, et al. Causality constrained machine learning framework enhances the reliability and spatiotemporal generalization in ecosystem respiration estimation[J]. *Agricultural and Forest Meteorology*, 2025, 372: 110718.
- [25] SIAMI-NAMINI S, TAVAKOLI N, SIAMI NAMIN A. A comparison of ARIMA and LSTM in forecasting time series[C]// 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). Piscataway, New Jersey, USA: IEEE, 2019: 1394-1401.
- [26] 冯海宽, 樊意广, 陶惠林, 等. 利用无人机高光谱影像的冬小麦氮含量监测[J]. *光谱学与光谱分析*, 2023, 43(10): 3239-3246.
FENG H K, FAN Y G, TAO H L, et al. Monitoring of winter wheat nitrogen content using UAV-borne hyperspectral Imagery[J]. *Spectroscopy and Spectral Analysis*, 2023, 43(10): 3239-3246.
- [27] 王超, 王建明, 冯美臣, 等. 基于多变量统计分析的冬小麦长势高光谱估算研究[J]. *光谱学与光谱分析*, 2018, 38(5): 1520-1525.
WANG C, WANG J M, FENG M C, et al. Hyperspectral estimation on growth status of winter wheat by using the multivariate statistical analysis[J]. *Spectroscopy and Spectral Analysis*, 2018, 38(5): 1520-1525.
- [28] 罗琦, 茹晓雅, 姜元, 等. 基于机器学习与气象灾害指标的苹果相对气象产量预测[J]. *农业机械学报*, 2023, 54(9): 352-364.
LUO Q, RU X Y, JIANG Y, et al. Prediction of apple relative meteorological yield based on machine learning and meteorological disaster indices[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2023, 54(9): 352-364.
- [29] 赵泽阳, 李美玲, 徐伟, 等. 基于无人机多时相特征的冬小麦产量预测模型研究[J]. *麦类作物学报*, 2025, 45(8): 1089-1100.
ZHAO Z Y, LI M L, XU W, et al. Yield prediction model of winter wheat based UAV-multi-temporal and multi-feature[J]. *Journal of Triticeae Crops*, 2025, 45(8): 1089-1100.
- [30] 王来刚, 郑国清, 郭燕, 等. 融合多源时空数据的冬小麦产量预测模型研究[J]. *农业机械学报*, 2022, 53(1): 198-204, 458.
WANG L G, ZHENG G Q, GUO Y, et al. A study on a winter wheat yield prediction model integrating multi-source spatiotemporal Data[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2022, 53(1): 198-204, 458.

Field Maize Yield Prediction Model Based on Causal Inference and Machine Learning in Agricultural Fields

WANG Yi^{1,6}, CUI Xitong¹, WANG Chen¹, XIONG Baowei¹, SHAO Guomin²,

WANG Wanying¹, CAO Pei^{3*}, HAN Wenting^{4,5*}

(1. School of Information Science and Engineering, Xi'an University of Finance and Economics, Xi'an 710100, China; 2. State Key Laboratory of Eco-hydraulics in Northwest Arid Region, Xi'an University of Technology, Xi'an 710048, China; 3. College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling 712100, China; 4. Institute of Water-saving Agriculture in Arid Areas (IWSA), Northwest A&F University, Yangling 712100, China; 5. National Engineering Laboratory for Crop High-efficiency Water Use, Yangling 712100, China; 6. Intelligent Financial Collaborative Trusted Computing Key Laboratory of Shaanxi Province for Higher Education Institutions, Xi'an 710100, China)

Abstract: [Objective] Maize is one of the most important staple crops in the world and serves as a cornerstone of food security and agricultural sustainability. Accurate and timely prediction of maize yield is essential for optimizing agricultural management practices, supporting market regulation, and guiding policy decisions related to food supply and climate adaptation. In recent years, data-driven yield prediction methods based on machine learning and deep learning have achieved notable improvements in predictive accuracy. However, most existing approaches primarily rely on statistical correlations among variables and often treat influencing factors as independent predictors, without explicitly addressing the complex causal mechanisms and time-lagged interactions that govern crop growth processes. This limitation may lead to reduced model interpretability and compromised robustness under changing environmental conditions. To address these challenges, a novel maize yield prediction framework that integrates causal inference with a hybrid deep learning model was proposed, aiming to improve both predictive performance and mechanistic understanding. **[Methods]** Multi-source heterogeneous datasets collected across the maize growing season were utilized, including remote sensing-derived vegetation indices, meteorological variables (such as temperature and precipitation), soil profile moisture measurements at multiple depths, and crop observation data corresponding to key phenological stages. First, the Peter-Clark and momentary conditional independence (PCMCI) causal discovery algorithm was applied to systematically identify causal relationships between maize yield and its potential driving factors. The PCMCI method enables the detection of both contemporaneous and time-lagged causal links while effectively controlling for confounding effects in high-dimensional time series data. Through this process, the causal structure of yield formation was explicitly characterized, and key variables with statistically significant causal impacts were selected as inputs for the prediction model. Subsequently, a hybrid moving average, convolutional neural network-long short-term memory (MA-CNN-LSTM) model was constructed to capture the complex spatiotemporal patterns in the causally screened input variables. Specifically, a moving average module was employed as a preprocessing step to suppress high-frequency noise and enhance signal stability. A CNN was then used to extract latent correlation features among multiple variables, reflecting their joint influence on yield formation. Finally, an LSTM network was adopted to model temporal dependencies and cumulative effects across the growing season, enabling effective representation of dynamic yield responses. **[Results and Discussions]** The causal analysis revealed that soil moisture at depths of 10 cm and 50 cm exerted a significant positive influence on maize yield ($P < 0.01$), with deeper soil moisture showing a stronger and more persistent time-lagged effect. This finding highlighted the critical role of subsurface water availability in sustaining crop growth during later developmental stages. In addition, vegetation indices such as the modified chlorophyll absorption ratio index and the normalized difference vegetation index exhibited significant short-term causal relationships with yield during the mid-growth stage of maize, indicating their sensitivity to canopy structure and photosynthetic activity during this period. Comparative experiments conducted against traditional statistical models and conventional machine learning approaches demonstrated that the proposed PCMCI-MA-CNN-LSTM framework consistently achieved superior predictive performance. On the test dataset, the coefficient of determination (R^2) reached 0.955, while the mean absolute error (MAE) and root mean square error (RMSE) were reduced to 1.201 kg/mu and 1.474 kg/mu (1 hm²=15 mu). These results indicated that incorporating causal variable selection effectively enhances model accuracy and stability by reducing redundant and spurious correlations. **[Conclusions]** The results confirm that incorporating causal analysis into yield modeling provides a robust basis for identifying key driving variables and effectively enhances the accuracy and interpretability of maize yield prediction. The proposed framework offers a promising approach for precision agriculture and decision support in crop yield forecasting, particularly under complex and dynamic agro-environmental conditions.

Keywords: peter-clark and momentary conditional independence (PCMCI); moving average; convolutional neural network; long short-term memory; yield prediction; machine learning; causal inference

Foundation items: National Social Science Fund Project (23BGL252); Basic Research Plan Project for Natural Sciences of Shaanxi Province (2022JQ-363); Key Innovation Chain Projects of Shaanxi Province(2024NC-ZDCYL-05-01); Key Innovation Chain Projects of Shaanxi Province (2023-ZDLNY-58); Xi'an Municipal Science and Technology Plan Project (25NJSYB00014)

First author: WANG Yi, E-mail: wang_yi@xaufe.edu.cn

***Corresponding author:** 1. CAO Pei, E-mail: caopei2020@nwafu.edu.cn; 2. HAN Wenting, E-mail: hanwt2000@126.com