DOI：10.13733/j.jcam.issn.2095-5553.2024.10.030
贾金豹，朱成娟，汪家全，等．基于机器学习结合多因子组合的玉米估产研究[J]．中国农机化学报，2024，45(10)：206－214
Jia Jinbao，Zhu Chengjuan，Wang Jiaquan，et al. Research on maize yield estimation based on machine learning combined with multi-factor combination [J]. Journal of Chinese Agricultural Mechanization，2024，45(10)：206－214

基于机器学习结合多因子组合的玉米估产研究*

贾金豹[1]，朱成娟[2]，汪家全[1]，周鹏[3]

（1. 信阳艺术职业学院，河南信阳，464000；2. 大连交通大学交通运输工程学院，大连市，116028；
3. 河南农业大学信息与管理科学学院，郑州市，450003）

摘要：玉米作为河南省主要种植作物之一，作物产量预测对区域贸易和粮食安全具有重要意义。为建立简单、及时、准确的作物叶面积指数*LAI*和产量预测模型，采用多元线性回归MLR、偏最小二乘回归PLSR和决策树DT机器学习技术，结合玉米生理参数因子(*P*1)、光谱特征波段(*P*2)、土壤性质参数(*P*3)和气象参数(*P*4)进行多因子组合构建玉米*LAI*和产量的估测模型。研究结果表明，在3种机器学习方法中，籽粒形成期的*LAI*估测精度显著优于其他生育时期，而成熟期的产量模型估测精度显著优于其他时期；在5种多因子组合中，PLSR算法结合*P*1＋*P*2＋*P*3＋*P*4多因子组合构建的模型达到最高精度，其中*LAI*估测最高为R_v^2＝0.84，$RMSE_v$＝0.38，产量估测最高为R_v^2＝0.79，$RMSE_v$＝982 kg/hm^2。为我国北方玉米种植区的玉米生长和产量预测提供技术支持和理论依据，提高预测的准确性和效率，对农业生产管理和决策制定具有重要意义。

关键词：玉米；机器学习；高光谱；生理指标；叶面积指数；产量

中图分类号：S513　　**文献标识码：**A　　**文章编号：**2095-5553 (2024) 10-0206-09

Research on maize yield estimation based on machine learning combined with multi-factor combination

Jia Jinbao[1], Zhu Chengjuan[2], Wang Jiaquan[1], Zhou Peng[3]

(1. *Xinyang Vocational College of Aart*, *Xinyang*, 464000, *China*; 2. *School of Traffic and Transportation Engineering*, *Dalian Jiaotong University*, *Dalian*, 116028, *China*; 3. *College of Information and Management Science*, *Henan Agricultural University*, *Zhengzhou*, 450003, *China*)

Abstract: As one of the main crops in Henan Province, maize yield prediction holds significant importance for regional trade and food security. In order to establish a simple, timely, and accurate model for predicting crop *LAI* and yield, this study employs multiple linear regression (MLR), partial least squares regression (PLSR), and decision tree (DT) machine learning techniques. These techniques are combined with multi-factor data, including maize physiological parameters (*P*1), spectral characteristic bands (*P*2), soil property parameters (*P*3), and meteorological parameters (*P*4), to construct estimation models for maize *LAI* and yield. The study results indicate that among the three machine learning methods, the *LAI* estimation accuracy during the grain filling stage is significantly higher than in other growth stages, while the yield estimation accuracy during the maturity stage is significantly higher than in other stages. Among the five multi-factor combinations, the PLSR algorithm combined with the *P*1＋*P*2＋*P*3＋*P*4 multi-factor combination has achieved the highest accuracy, with the highest *LAI* estimation at R_v^2＝0.84 and $RMSE_v$ = 0.38, and the highest yield estimation at R_v^2＝0.79 and $RMSE_v$ = 982 kg/hm^2. These findings provide technical support and theoretical basis for regional maize growth and yield prediction in the maize-growing areas of northern China, enhancing prediction accuracy and efficiency, and are of great significance for agricultural production management and decision-making.

Keywords: maize; machine learning; hyperspectral; physiological indicators; LAI; yield

收稿日期：2024年3月11日　　修回日期：2024年6月24日
*基金项目：河南省高等学校重点科研项目(23B413004)
第一作者：贾金豹，男，1979年生，河南信阳人，讲师；研究方向为农业信息技术和算法。E-mail：jiajinbao1213@163.com

中国知网 https://www.cnki.net

0　引言

玉米(maize)作为一种重要的粮食、饲料和生物能源作物，在世界范围内具有广泛的种植面积和消费群体。河南省作为我国玉米的主产区之一，其玉米的生长和产量直接影响着我国农业生产的效益和安全。因此，深入研究玉米生长过程对于探索和预测玉米叶面积指数和产量具有重要的理论和实践意义。

针对玉米生物量监测的传统研究方法主要通过人工采样，但此方法耗时费力，且无法满足大面积玉米生物量实时、精确监测的需求[1, 2]。目前，通过经验统计模型和基于机器学习算法估测模型来改进作物产量预测越来越来受人们重视[3]。传统的统计模型通过建立天气变量(温度、降水、太阳辐射等)与作物田间生理参数变量，包括：株高、SPAD(Soil and Plant Analyzer Development，一种衡量植物叶片中叶绿素相对含量的指标)、叶绿素含量、鲜重和干重等)之间的回归方程来预测产量，并在不同时间和空间尺度上进行预测[4]。这种回归结果清楚地显示了气候因子或作物生理参数等单一因子对产量的影响，但它们因相对较低的解释能力而存在较大争议，而控制产量的主要因素往往是多种影响因子共同决定的，且随着生长阶段的变化而变化[5]。因此，这些单一因子构建估测模型的精度往往达到一定的数值后较难再有提升，很难应用于更大的区域[6]。

机器学习的广泛应用已经证明了其在数据挖掘和农业分析中的强大性能，从而更有效地进行作物生长监测和产量预测[7]。玉米生长受到土壤条件、气候和各种变量因子的影响，这些因子之间的相互作用对作物产量的形成起着重要作用[8]。然而，在许多研究中选择的变量都是基于整个生长季节的生理参数或者单一影响因子，这意味着最终的产量直到收获时才能被估计出来[9]。根据研究调查可知，多因子组合同时作为变量结合机器学习估测玉米产量的研究尚少。确定最佳的多因子组合，可以更好地融合多因子对玉米产量的估测优势，具有提高玉米产量预测模型精度的潜力[10]。

随着科技的进步和研究方法的创新，多因子结合机器学习成为了当前作物生长研究的热点领域之一[11]。其中，部分研究主要集中在气候因子和栽培管理策略对玉米生长的影响，如温度、光照、水分、施肥等[12]；同时，机器学习方法也被应用于玉米生长模型的构建和预测[13]。另外，研究以农业区域化特征为主，主要关注玉米的栽培管理、根系发育和产量等方面，但在机器学习的应用方面仍然相对较少[3, 14]。越来越多的学者基于利用遥感数据、气象数据、土壤信息等多源数据进行作物产量估测，并运用机器学习方法构建预测模型，取得了较好的结果。彭慧文[15]、刘帅兵[16]等利用气候参数和土壤环境因子的相互作用，结合机器学习和作物生长模型对玉米地上生物量和产量的模拟结果得到有效改善。吴永清等[17]采用MLR和PLSR算法对小麦、玉米等谷物的产量进行估测，系统的研究了不同算法预测玉米产量方面的能力。

为了优化当前玉米产量预测的准确性与时效性不足的问题。本研究旨在利用机器学习技术结合多因子组合，构建及时、准确的玉米叶面积指数和产量的预测模型。整合玉米的生理参数因子($P1$)、光谱特征波段($P2$)、土壤性质参数($P3$)和气象参数($P4$)等多源数据，探究参数因子对玉米LAI和产量拟合结果的影响；探索多元线性回归(MLR)、偏最小二乘回归(PLSR)和决策树(DT)算法在玉米生长和产量预测中的适用性，并分析不同生长时期对玉米叶面积指数和产量预测的影响程度；探寻多种机器学习算法结合多因子组合的最优玉米LAI和产量预测模型。

1　材料和方法

1.1　研究区概况

研究区位于我国华北平原河南省许昌市(113°54′E，33°57′N)。农作物种类主要为玉米和冬小麦，耕作分为夏、秋两季。许昌市属暖温带大陆性季风气候，降水集中于夏末秋初，多年平均降水量约为700 mm，平均气温15.1 ℃，雨热同季，光照时间充足，农业发展地理条件优越。研究区种植45个玉米小区，每个小区长8 m，宽5 m，且采用相同的水肥管理，试验期间采集了2022年4—9月的玉米生理指标参数和其他各类数据，图1为研究区的概况图。

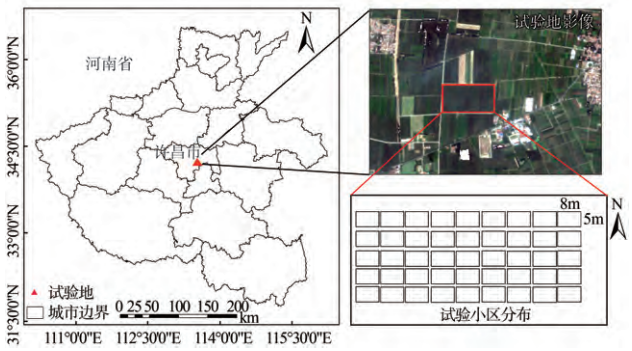


图 1　研究区概况图

Fig. 1 Overview of the study area

1.2　多源数据收集

1.2.1　生理指标参数

测定多个玉米生长发育指标参数，主要测定玉米的指标参数有株高、$SPAD$(叶绿素相对含量)[18]、叶绿素含

中国知网　https://www.cnki.net

量、鲜重、干重和叶面积指数[19]，采集的玉米生育时期分别为苗期(S1)、抽雄期(S2)、籽粒形成期(S3)和成熟期(S4)，将以上指标统一为一类影响因子，即生理指标参数($P1$)，最后在玉米收获期测定每个小区的产量。

1.2.2 光谱参数及特征选取

玉米冠层光谱采用FieldSpec 4地物光谱仪测定，光谱范围为350～2 500 nm，350～1 000 nm和1 000～2 500 nm的光谱分辨率分别为3 nm和10 nm。本研究是在玉米的四个关键生育时期(中午11:00—13:00)利用光谱仪垂直于地面照射玉米冠层获得的光谱反射率数据，每个小区采集5条光谱数据，经过S－G(Savitzky－Golay)平滑预处理[20]，并取平均值，再利用连续投影算法(Successive Projections Algorithm，SPA)筛选出对玉米冠层响应良好的特征波段反射率参数($P2$)。

1.2.3 土壤参数

土壤的特性对植物的生长发育至关重要，并对作物产量有重大影响，本研究的样本在同一个试验田中，从整个试验田中在玉米播种之前随机选取5个点的耕层土(0～20 cm)的平均值作为本试验田的土壤参数来源。获取的土壤物理和化学参数($P3$)主要包括土壤pH、土壤容重、有机碳含量、总氮含量、速效磷、有效钾、物理砂性和物理粘性(表1)，各项指标均测定3次并取平均值，详细的指标测定方法见参考文献[21-23]。

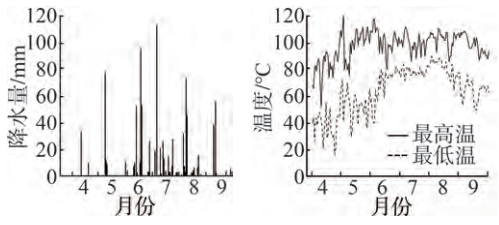
表 1 研究区土壤测定参数

Tab. 1 Soil determination parameters in the study area

参数	数值	参数	数值
pH	6.9	速效磷/(mg·kg^{-1})	9.4
土壤容重/(g·cm^{-3})	1.37	有效钾/(mg·kg^{-1})	127.07
有机碳/%	1.79	物理砂性/%	70.8
总氮/%	2.05	物理粘性/%	29.2

1.2.4 气象参数

气象参数是从国家气象科学中心获取，该数据为2022年4—9月河南省许昌市气象站点的逐日气象数据($P4$)，覆盖了玉米生长发育的整个周期，本研究使用的主要气象变量包括最高温、最低温和降水量(图2)。



(a) 降水量 (b) 最高温、最低温

图 2 研究区气象参数

Fig. 2 Meteorological parameters of the study area

1.3 研究方法

本文按照图3所示的步骤方法建立研究区2022年的玉米产量估算模型，总共45个种植小区，随机将小区分为两部分，即建模集($n=30$)和验证集($n=15$)。

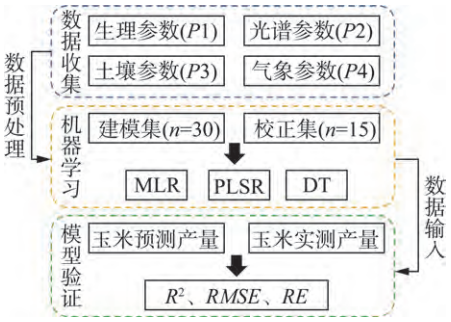


图 3 技术路线

Fig. 3 Technology road map

本研究采用S－G平滑[23]与连续投影算法(Successive Projections Algorithm，SPA)，对光谱数据进行预处理以及特征光谱的筛选[24]，从而达到降维、提取重要特征、去除冗余信息和可视化光谱特征分布等要求。随后，采用多元线性回归(Multiple Linear Regression，MLR)[24]、偏最小二乘回归(Partial Least Squares Regression，PLSR)[24]和决策树(Decision Tree，DT)[25]，三种机器学习算法分别构建针对玉米产量的估测模型，其中建模集与校正集的比例为2∶1。

1.4 数据组合

考虑到多种影响因子的相互作用和实际因子之间的内在联系性，根据影响因子的重要程度组合出5组多因子组合，分别是单因子($P1$、$P2$)，双因子组合($P1+P2$、$P3+P4$)，多因子组合($P1+P2+P3+P4$)。选择$P1$和$P2$作为单因子是因为其在预测玉米LAI和产量模型中最具代表性和重要性。$P1+P2$和$P3+P4$的双因子组合则是考虑到数据类型的共线性且需要评估重要因子间的相互作用和覆盖面。多因子组合($P1+P2+P3+P4$)则综合考虑所有重要因子及其交互作用，这种选择方法确保了在构建玉米LAI和产量模型时，能够全面、平衡地考虑各因子的影响，从而探索不同多因子组合对构建玉米LAI和产量的模型精度影响。

本文在运用三种机器学习模型进行拟合时，参数选择是通过交叉验证方法确定的，以确保模型的最佳性能和稳定性。

1.5 模型评估

为了评价玉米产量估计的准确性，研究采用决定系数R^2、均方根误差$RMSE$和相对误差RE作为定量指标。其中$RMSE$是预测值与真实观测值之间的差异的平方和的平均值的平方根，它衡量了模型的预测误差的大小，数值越小表示预测精度越高。RE是用来衡量预测值与真实观测值之间的相对差异。R^2的取值范围在0～1之间，越接近1表示模型能更好地解

中国知网 https://www.cnki.net

释因变量的变异性，而越接近 0 表示模型的解释能力较低。定量指标的计算方法如式(1)～式(3)所示。

$$R^2=1-\frac{\sum_{i=1}^{n}\left(y_i-o_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2} \tag{1}$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-o_i\right)^2}{n}} \tag{2}$$

$$RE=\frac{|y_i-o_i|}{y_i} \tag{3}$$

式中：n ——训练或者验证集的样本个数；

$\bar{y}$ ——实测平均值；

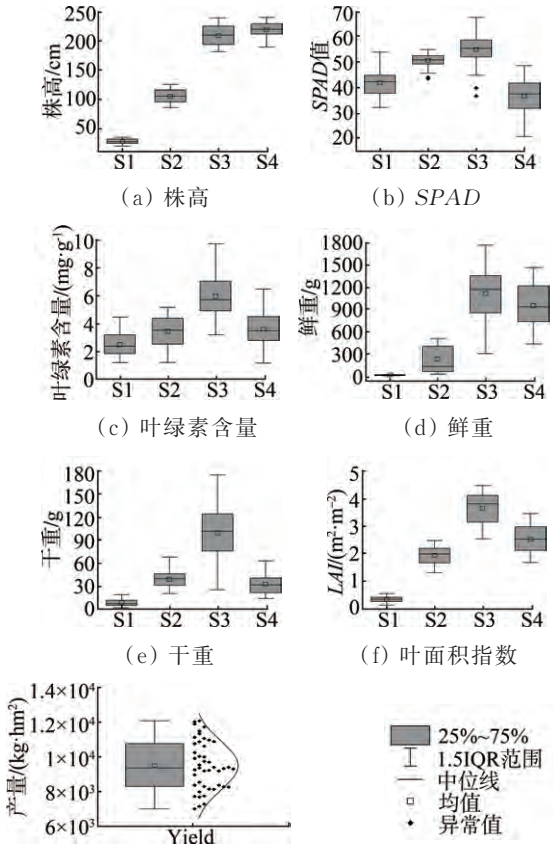
y_i ——第 i 个实测值；

o_i ——预测值[26]。

2　结果与分析

2.1　玉米生长期生理指标表现

选取玉米的 4 个关键生育时期的多个生理指标组合作为玉米产量估测的一个影响因子，记为生理参数($P1$)，获取的生理参数分布范围如图 4 所示。



(a) 株高　(b) *SPAD*　(c) 叶绿素含量　(d) 鲜重　(e) 干重　(f) 叶面积指数　(g) 产量

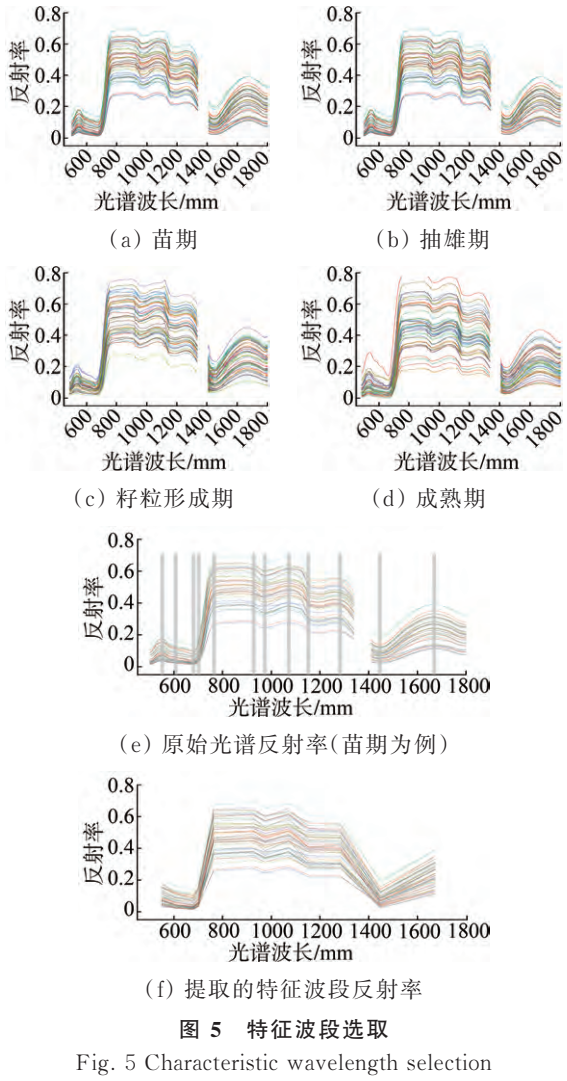
图 4　玉米生长指标参数

Fig. 4 Growth index parameters of maize

由图 4 可知，S1、S2、S3、S4 分别对应的生育时期为苗期、抽雄期、籽粒形成期和成熟期。株高随着玉米生长呈现先增高后平稳的趋势，其他 5 个生理指标(*SPAD*、叶绿素含量、鲜重、干重和 *LAI*)均呈现先增高后在成熟期有略微下降的趋势，这个趋势是符合玉米各项生理指标在生育时期的规律。

2.2　特征波段的选取

利用 ASD 光谱仪获取了玉米冠层连续非成像的光谱反射率数据，由于光谱波段之间的波长相近，使得其光谱反射率的共线性较强，表征的光谱信息存在较多的冗余，所以本文通过连续投影算法(Successive Projections Algorithm，SPA)[27]，减少光谱数据的共线性，筛选出有效的特征波段，通过连续投影算法筛选的结果如图 5 所示。提取的特征波段如表 2 所示。



(a) 苗期　(b) 抽雄期　(c) 籽粒形成期　(d) 成熟期　(e) 原始光谱反射率(苗期为例)　(f) 提取的特征波段反射率

图 5　特征波段选取

Fig. 5 Characteristic wavelength selection

表 2　特征波段的选取

Tab. 2 Characteristic wavelength selection

光谱预处理	特征波长筛选方法	特征中心波长/nm
S－G 平滑	SPA	550、605、678、700、763、926
S－G		972、1 070、1 150、1 281、1 445、1 669

中国知网　https://www.cnki.net

其中，光谱反射率在波长1 340～1 410 nm空白的原因是这个范围内光谱反射率受到水汽和环境噪声的干扰，导致光谱信号反射率发生无规律的上下波动，对于特征波长的提取无意义，且影响呈现效果，故去除了该范围的波长反射率。最终利用SPA算法，获取12个关键光谱特征中心波段（图5(e)和5(f)），获得的特征波段是对于玉米生长发育有较强的光谱响应，是监测玉米生长发育的重要光谱波段，包括可见光波段和近红外波段。

2.3 机器学习估测玉米*LAI*和产量

2.3.1 估测不同生育时期玉米*LAI*和产量

利用三种机器学习算法（MLR，PLSR和DT），结合玉米生理指标参数（*P*1），构建玉米*LAI*和产量的估测模型，如表3所示。

表 3 利用机器学习算法估算玉米S1~S4的*LAI*和产量

Tab. 3 *LAI* and yield of maize S1－S4 were estimated using a machine learning algorithm

机器学习	生育时期	*LAI*/($m^2 \cdot m^{-2}$)			产量/($kg \cdot hm^{-2}$)		
		R_v^2	$RMSE_v$	*RE*/%	R_v^2	$RMSE_v$	*RE*/%
MLR－*P*1	S1	0.23	0.97	17.7	0.22	2 624	19.7
	S2	0.45	0.82	14.9	0.42	2 136	14.9
	S3	0.70	0.45	8.2	0.62	1 541	13.2
	S4	0.66	0.51	9.3	0.68	1 349	11.5
PLSR－*P*1	S1	0.28	0.94	17.1	0.25	2 575	19.4
	S2	0.47	0.80	14.6	0.44	2 011	14.7
	S3	0.77	0.38	6.9	0.59	1 626	13.5
	S4	0.65	0.55	10.0	0.72	1 233	8.9
DT－*P*1	S1	0.24	0.96	17.5	0.21	2 640	20.0
	S2	0.47	0.80	14.6	0.43	2 035	14.8
	S3	0.71	0.43	7.8	0.60	1 651	13.4
	S4	0.65	0.54	9.8	0.65	1 428	12.7

注：S1～S4为玉米的生育时期，分别为苗期、抽雄期、籽粒形成期和成熟期；R_v^2为验证集的决定系数，$RMSE_v$为验证集的均方根误差，*RE*为样本的相对误差，下同。

由表3可知，三个模型算法对于玉米*LAI*的估测精度整体优于对于玉米产量的估产精度，研究表明，PLSR－*P*1模型在估测玉米*LAI*和产量方面的精度优于MLR－*P*1和DT－*P*1，其中在籽粒形成期（S3）估测*LAI*的精度达到最高（R_v^2=0.77，$RMSE_v$=0.38，*RE*=6.9%）；对于玉米产量的R_v^2，$RMSE_v$和*RE*分别达到了0.72、1 233 kg/hm^2和8.9%。此外，研究还发现MLR－*P*1模型估测精度的稳定性优于PLSR－*P*1和DT－*P*1。

2.3.2 多因子组合构建玉米*LAI*估测模型

使用三种机器学习方法结合玉米生理参数（*P*1）、光谱特征波段（*P*2）、土壤性质参数（*P*3）和气象参数（*P*4）4种因子数据，构建了玉米*LAI*估测模型，多因子组合的方法总共分为5种，分别为*P*1、*P*2、*P*1＋*P*2、*P*3＋*P*4和*P*1＋*P*2＋*P*3＋*P*4。选择四因素组合（*P*1＋*P*2＋*P*3＋*P*4）而不是三因素组合是考虑到四因素能够更全面地覆盖影响玉米生长和产量的各个方面，包括生理参数、光谱特征、土壤性质和气象参数，确保模型具有更高的信息完整性和代表性。同时，四因素组合能够更好地捕捉复杂的交互作用，提高模型的预测精度和稳定性，在前期验证支持的基础上，直接采用四因素组合也能节省研究资源并保证模型的泛化能力。

由表4可知，建模集的估测决定性系数优于校正集的决定性系数，且可以发现一般在建模方法不变的情况下，随着预测因子的增加玉米*LAI*的估测精度也随之增大，即*P*1和*P*2单个的决定系数略低于*P*1＋*P*2的决定系数。其中，PLSR模型结合*P*1＋*P*2因子组合模型对*LAI*的估测精度优于MLR和DT的建模精度，R_c^2和R_v^2分别达到0.84和0.81。预测因子达到4个时，各个模型的*LAI*估测精度均有略微提高，其中，利用PLSR结合*P*1＋*P*2＋*P*3＋*P*4预测因子的组合建模使得玉米*LAI*的估测精度达到最高，校正集R_v^2达到0.84，$RMSE_v$为0.38，*RE*为7.2%。

然而由表4进一步发现并非总是随着预测因子的增加估测精度一定会增大，其中，土壤性质参数和气象参数（*P*3＋*P*4）预测因子组合构建*LAI*估测模型的估测精度明显低于*P*1和*P*2的估测精度。表明，土壤性质参数与气象参数组合构建的*LAI*估测模型对于*LAI*的影响权重低于玉米的生理参数和光谱特征波段。

中国知网 https://www.cnki.net

表 4　多因子组合建模和校正估测玉米 *LAI*

Tab. 4 Multi-factor combination modeling and correction for estimation of maize *LAI*

机器学习	多因子组合	模型建立(训练集 $n=30$)		模型检验(校正集 $n=15$)		
		R_c^2	$RMSE_c$	R_v^2	$RMSE_v$	RE/%
MLR	$P1$	0.62	0.53	0.59	0.59	11.2
	$P2$	0.58	0.59	0.52	0.65	12.3
	$P1+P2$	0.78	0.44	0.76	0.46	8.7
	$P3+P4$	0.32	0.77	0.29	0.83	15.7
	$P1+P2+P3+P4$	0.82	0.39	0.79	0.40	7.6
PLSR	$P1$	0.69	0.48	0.68	0.47	8.9
	$P2$	0.65	0.50	0.62	0.54	10.2
	$P1+P2$	0.84	0.38	0.81	0.39	7.4
	$P3+P4$	0.32	0.78	0.33	0.80	15.1
	$P1+P2+P3+P4$	0.85	0.37	0.84	0.38	7.2
DT	$P1$	0.58	0.59	0.54	0.39	7.4
	$P2$	0.56	0.60	0.49	0.58	11.0
	$P1+P2$	0.70	0.46	0.68	0.47	8.9
	$P3+P4$	0.25	0.84	0.23	0.85	16.1
	$P1+P2+P3+P4$	0.79	0.41	0.74	0.44	8.3

2.3.3　多因子组合构建玉米产量估测模型

利用三种建模方法结合玉米生理参数($P1$)、光谱特征波段($P2$)、土壤性质参数($P3$)和气象参数($P4$)因子数据，构建得到多因子组合的玉米产量估测模型，如表5所示。

表 5　多因子组合的建模和校正估测玉米产量

Tab. 5 Modeling and correction of multi-factor combinations to estimate maize yield

机器学习	多因子组合	模型建立(训练集 $n=30$)		模型检验(校正集 $n=15$)		
		R_c^2	$RMSE_c$/(kg·hm^{-2})	R_v^2	$RMSE_v$/(kg·hm^{-2})	RE/%
MLR	$P1$	0.66	1 510	0.64	1 422	12.5
	$P2$	0.59	1 727	0.55	1 786	14.1
	$P1+P2$	0.75	1 143	0.74	1 241	9.5
	$P3+P4$	0.22	2 657	0.19	2 743	21.1
	$P1+P2+P3+P4$	0.79	994	0.77	1 098	8.2
PLSR	$P1$	0.70	1 388	0.68	1 299	11.5
	$P2$	0.67	1 489	0.64	1 520	12.4
	$P1+P2$	0.77	1 067	0.73	1 240	9.9
	$P3+P4$	0.23	2 561	0.23	2 624	19.6
	$P1+P2+P3+P4$	0.82	915	0.79	982	7.7
DT	$P1$	0.66	1 502	0.61	1 537	13.2
	$P2$	0.61	1 695	0.60	1 603	13.4
	$P1+P2$	0.73	1 267	0.70	1 350	10.9
	$P3+P4$	0.20	2 644	0.19	2 683	21.1
	$P1+P2+P3+P4$	0.77	1 051	0.72	1 270	10.1

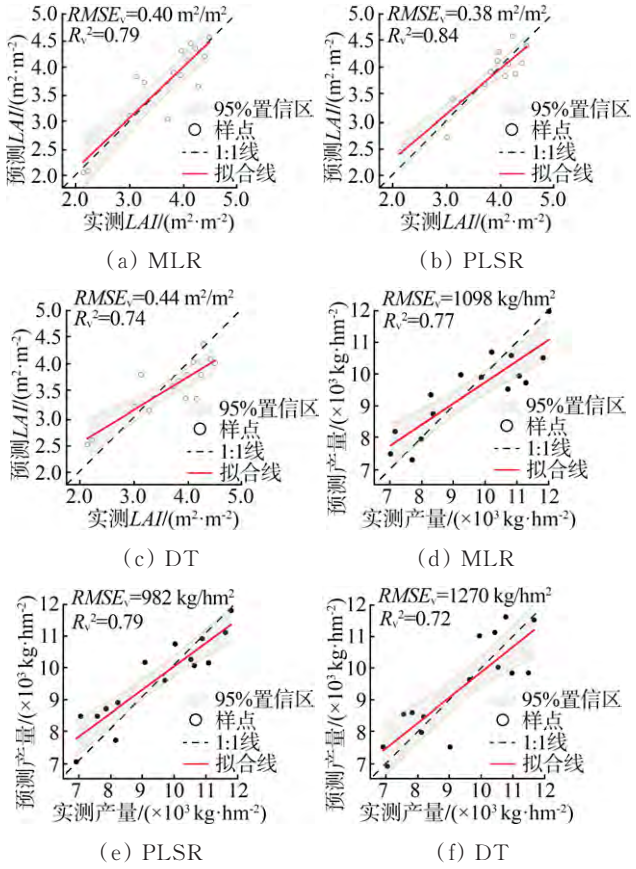
建模集的R_c^2优于校正集的R_v^2，在相同建模方法的情况下，随着预测因子的增加，玉米产量的估测精度也增大。单独考虑$P1$和$P2$的决定系数略低于考虑$P1+P2$的决定系数。且通过结合$P1+P2$因子的PLSR模型的估测精度优于MLR和DT模型的建模精度，其中R_c^2和R_v^2分别达到0.77和0.73。当预测因子增加到4个时，各模型的产量估测精度略微提高，其中利用PLSR结合$P1+P2+P3+P4$预测因子的组合模型使得玉米产量的估测精度最高，校正集R_v^2达到0.79，$RMSE_v$为982 kg/hm^2，RE为7.7%。然而，从表5可知，估测精度并非总是随着预测因子的增加就一定增大。特别是土壤性质参数和气象参数($P3+P4$)的预测因子组合建模所得到的产量估测模型的精度明显低于$P1$和$P2$的估测精度。这说明对玉米产量的预测中土壤性质参数和气象参数对产量的影响

中国知网　https://www.cnki.net

权重显著低于玉米的生理参数和光谱特征波段。

2.3.4 最优建模结果

根据表4和表5得到的三种机器学习结合多因子组合构建的玉米*LAI*和产量估测模型，研究结果显示MLR、PLSR和DT结合*P*1+*P*2+*P*3+*P*4多因子组合的模型估测精度优于其他4种多因子组合。图6为三种机器学习对玉米*LAI*和产量分别构建的最优建模拟合图。



(a) MLR (b) PLSR (c) DT (d) MLR (e) PLSR (f) DT

图 6 最优建模验证1∶1拟合图

Fig. 6 Optimal modeling validation of a 1∶1 fit

由图6可知，三种机器学习算法整体对于玉米*LAI*的估测精度优于对玉米产量的估测精度，其中，利用PLSR算法结合*P*1+*P*2+*P*3+*P*4多因子组合的估测精度达到最高(R_v^2=0.84，$RMSE_v$=0.38)。除此之外，PLSR建模方法对于玉米产量的估测精度(R_v^2=0.79，$RMSE_v$=982 kg/hm²)同样优于MLR和DT的玉米产量估测精度。研究结果表明，通过MLR与DT构建的玉米*LAI*和产量模型，在验证图中，预测值和实测值样点分布较为分散，有近1/4的验证样点分布在95%置信区间外，这样的结果会降低模型模拟的精度，使得估测精度低于验证样点集中于置信区间的PLSR建模方法。

3 讨论

3.1 机器学习在各生育时期的表现

玉米的关键生育时期包括苗期、茎叶生长期、抽雄期、籽粒形成期和成熟期，各个时期的*LAI*和产量对于玉米自身的生理指标参数(*P*1)响应程度差异较大。本文选取三种机器学习方法结合生理参数因子，构建了四个关键生育时期的*LAI*和产量估测模型，研究结果表明，玉米生长发育前期(苗期和抽雄期)机器学习对于*LAI*和生物量的估测效果较差，而在籽粒形成期，*LAI*的估测精度最高。贺佳等[28]研究发现，利用无人机光谱数据构建的植被指数在玉米抽雄期与成熟期之间估测精度较高，结果表明，所构建的NDRE指数的决定系数达到了0.75，表明玉米籽粒形成期是估测玉米*LAI*较为可靠的生育时期。在构建玉米产量估测模型时，苗期、抽雄期和籽粒形成期的估测精度低于机器学习在玉米成熟期产量估测精度。本文研究结果的原因可能在于，玉米生长发育前期的各项生理指标无法直接反映产量信息，且在生育前期表现出的指标特性与产量影响权重较小[29]。

3.2 多因子选取和组合对估测的影响

为了提高估测的准确性，本文采用三种机器学习方法，并结合多个因子的选取和组合，对玉米的*LAI*和产量构建模型并对其进行估测。在估测过程中，多个预测因子会对结果产生影响。这些因子包括植株生理参数(*P*1)、冠层特征波段(*P*2)、土壤性质(*P*3)和气象参数(*P*4)。组成单影响因子(*P*1、*P*2)、双影响因子(*P*1+*P*2、*P*3+*P*4)和多影响因子(*P*1+*P*2+*P*3+*P*4)共5种多因子组合。不同因子之间的组合对估测结果产生影响，机器学习算法通过多个因子之间的相互作用和影响，以更好地捕捉因子之间的复杂关系。例如，土壤性质参数和气温可能存在交互作用，对玉米的生长影响会相互叠加[30]。通过机器学习，可以利用这样的组合关系，提高估测的准确性。总所周知，在研究方法确定的情况下，随着添加的因子增加估测指标的精度也会随之增大，本研究在一定范围内是符合这项规律的。然而本研究结果也出现具有反差性的结果，由表4表明*P*3+*P*4多因子组合获得的决定系数明显低于*P*1和*P*2单因子估测玉米*LAI*和产量的决定系数。出现此结果的原因可能是土壤性质参数(*P*3)和气象参数(*P*4)是间接影响玉米生长发育的条件因子，它们之间不是直接关系，中间可能还存在转换条件，例如，土壤中的pH值最先是对玉米的根系生长环境产生影响，影响到根系蛋白质活性，进而间接对玉米地上部*LAI*和籽粒形成产生影响[31]。

3.3 机器学习结合多因子组合的表现

不同机器学习方法有其各自的适用条件和要求，本文分别采用MLR、PLSR和DT三种机器学习方法结合多因子组合对玉米的*LAI*和产量进行估测研究。

中国知网 https://www.cnki.net

在预测因子数目一致的情况下，PLSR算法构建的玉米*LAI*和产量估测模型精度优于MLR和DT估测精度。表明在同等数据条件下，PLSR算法能够充分利用输入特征之间的相关性，并且通过对输入特征因子进行线性组合，能够同时考虑多个输入特征之间的相互作用，从而具有更强的表达能力[32]。张亚倩[33]、谭先明[34]等研究发现，分别采用机器学习算法结合激光雷达和高光谱参数构建得到玉米*LAI*和产量估测精度较高的模型，结果表明，采用PLSR算法构建玉米*LAI*和产量估测精度达到最高，R^2分别为0.88和0.51，通过PLSR建立的预测模型，可以更好地估计玉米产量，为间作玉米的田间管理和生长监测提供理论和技术参考。欧阳玲等[35]基于NDVI、EVI和GNDVI构建的MLR为玉米产量估算最优模型（R^2=0.82，$RMSE$=1 354.5 kg/hm^2），精度达到了80.55%，为精准农业的发展提供了参考。

在未来的研究中，可以进一步探索其他机器学习算法，如随机森林（RF）、支持向量机（SVM）、反向传播神经网络（BPNN）、遗传算法（GA－BP）神经网络[36]等，以提高模型的预测精度。此外，还可以考虑引入更多的因子和特征参数，进一步提高模型的表达能力和泛化能力。综上所述，结合多因子组合的机器学习方法在玉米*LAI*和产量估测中具有巨大的潜力，可以为农业生产提供精准的决策支持。

4　结论

1）通过SPA筛选得到玉米冠层光谱的12个特征波段（*P*2），一定程度上减少了光谱数据的冗余性，且在利用PLSR算法构建的玉米*LAI*和产量模型中模型估测效果有较好的表现（*LAI*：R_v^2=0.62，$RMSE_v$=0.54；产量：R_v^2=0.64，$RMSE_v$=1 520 kg/hm^2）。

2）利用三种机器学习算法结合玉米生理参数因子（*P*1）构建玉米*LAI*和产量的估测模型中，籽粒形成期生理参数构建的*LAI*估测精度显著优于苗期、抽雄期和成熟期，其中，采用PLSR算法结合*P*1估测精度达到最高（PLSR＋*P*1，R_v^2=0.77，$RMSE_v$=0.38）。

3）成熟期的生理参数构建的产量估测精度显著优于其他三个生育时期，其中，采用PLSR算法结合*P*1估测精度达到最高（PLSR＋*P*1，R_v^2=0.72，$RMSE_v$=1 233 kg/hm^2）。采用4种影响因子组成5种多因子组合结合机器学习算法构建的玉米*LAI*和产量估测模型中，PLSR算法结合*P*1＋*P*2＋*P*3＋*P*4多因子组合构建的估测模型精度达到最高，其中玉米*LAI*估测最高为R_v^2=0.84，$RMSE_v$=0.38，玉米产量估测精度最高为R_v^2=0.79，$RMSE_v$=982 kg/hm^2。

这项研究的结果为中国北方玉米种植区的区域性玉米生理指标和产量的预测提供技术支持和理论依据。此外，通过整合多源数据结合利用机器学习方法，可以提高玉米产量预测的准确性和效率，对于农业生产管理和决策制定具有重要意义。

参　考　文　献

[1] 陈上，窦子荷，蒋腾聪，等. 基于聚类法筛选历史相似气象数据的玉米产量DSSAT－CERES－Maize预测[J]. 农业工程学报，2017，33(19)：147－155.
Chen Shang，Dou Zihe，Jiang Tencong，et al. Maize yield forecast with DSSAT－CERES－Maize model driven by historical meteorological data of analogue years by clustering algorithm [J]. Transactions of the Chinese Society of Agricultural Engineering，2017，33(19)：147－155.

[2] 竞霞，邹琴，白宗璠，等. 基于反射光谱和叶绿素荧光数据的作物病害遥感监测研究进展[J]. 作物学报，2021，47(11)：2067－2079.

[3] 岑海燕，朱月明，孙大伟，等. 深度学习在植物表型研究中的应用现状与展望[J]. 农业工程学报，2020，36(9)：1－16.
Cen Haiyan，Zhu Yueming，Sun Dawei，et al. Current status and future perspective of the application of deep learning in plant phenotype research [J]. Transactions of the Chinese Society of Agricultural Engineering，2020，36(9)：1－16.

[4] Cai Y，Guan K，Lobell D，et al. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches [J]. Agricultural and forest meteorology，2019，274：144－159.

[5] Filippi P，Jones J E，Wimalathunge S N，et al. An approach to forecast grain crop yield using multi-layered，multi-farm data sets and machine learning [J]. Precision Agriculture，2019，20(5)：1015－1029.

[6] 张颖，赵宽辽，路燕. 我国玉米生产要素贡献率和地区差异实证分析——基于21个玉米主产省（区、市）的面板数据[J]. 河南农业科学，2013，42(8)：182－185.

[7] 李静，陈桂芬，安宇. 基于优化卷积神经网络的玉米螟虫害图像识别[J]. 华南农业大学学报，2020，41(3)：110－116.

[8] Crane-Droesch A. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture [J]. Environmental Research Letters，2018，13(11)：114003.

[9] Chen Y，Zhang Z，Tao F. Improving regional winter wheat yield estimation through assimilation of phenology and leaf area index from remote sensing data [J]. European Journal of Agronomy，2018，101：163－173.

[10] 崔颖，蔺宏宏，谢云，等. AquaCrop模型在东北黑土区作物产量预测中的应用研究[J]. 作物学报，2021，47(1)：159－168.

[11] 王国栋，姜明，盛春蕾，等. 湿地生态学的研究进展与展望[J]. 中国科学基金，2022，36(3)：364－375.

中国知网　https://www.cnki.net

[12] 杨艳昭，杨玲，张伟科，等．西辽河流域玉米水分平衡时空分布格局[J]．干旱区资源与环境，2014，28(4)：147－152.

[13] Kuradusenge M，Hitimana E，Hanyurwimfura D，et al. Crop yield prediction using machine learning models：Case of Irish potato and maize [J]. Agriculture，2023，13(1)：225.

[14] 杜兆辉，和贤桃，杨丽，等．玉米精准变量播种技术与装备研究进展[J]．农业工程学报，2023，39(9)：1－16.
Du Zhaohui，He Xiantao，Yang Li，et al. Research progress on precision variable-rate seeding technology and equipment for maize [J]. Transactions of the Chinese Society of Agricultural Engineering，2023，39(9)：1－16.

[15] 彭慧文，赵俊芳，谢鸿飞，等．作物模型应用与遥感信息集成技术研究进展[J]．中国农业气象，2022，43(8)：644－656.

[16] 刘帅兵，杨贵军，景海涛，等．基于无人机数码影像的冬小麦氮含量反演[J]．农业工程学报，2019，35(11)：75－85.
Liu Shuaibing，Yang Guijun，Jing Haitao，et al. Retrieval of winter wheat nitrogen content based on UAV digital image [J]. Transactions of the Chinese Society of Agricultural Engineering，2019，35(11)：75－85.

[17] 吴永清，李明，张波，等．高光谱成像技术在谷物品质检测中的应用进展[J]．中国粮油学报，2021，36(5)：165－173.

[18] 马红雨，李仙岳，孙亚楠，等．基于无人机遥感的不同控释肥夏玉米SPAD差异性[J]．排灌机械工程学报，2023，41(12)：1261－1267.

[19] 郭占强，肖国举，李秀静，等．不同土壤有机碳含量对玉米光合生理及生长发育的影响[J]．干旱地区农业研究，2022，40(1)：238－246.

[20] 王玉娜，李粉玲，王伟东，等．基于无人机高光谱的冬小麦氮素营养监测[J]．农业工程学报，2020，36(22)：31－39.
Wang Yuna，Li Fenling，Wang Weidong，et al. Monitoring of winter wheat nitrogen nutrition based on UAV hyperspectral images [J]. Transactions of the Chinese Society of Agricultural Engineering，2020，36(22)：31－39.

[21] 张孟豪，吴玲，陈静，等．蚯蚓对废纸屑再利用及养分贫瘠土壤综合质量的影响[J]．生态学报，2022，42(12)：5034－5044.

[22] 李百云，李慧，郭鑫年，等．基于最小数据集的宁夏耕地土壤质量评价[J]．江苏农业科学，2021，49(9)：195－201.

[23] 陈蒙蒙，兰玉彬，王国宾，等．基于土壤多参数监测系统的田间持水量试验研究[J]．中国农机化学报，2021，42(1)：130－135，244.
Chen Mengmeng，Lan Yubin，Wang Guobin，et al. Experimental study on field capacity based on soil multi-parameter monitoring system [J]. Journal of Chinese Agricultural Mechanization，2021，42(1)：130－135，244.

[24] 赵金龙，张学艺，李阳．机器学习算法在高光谱感知作物信息中的应用及展望[J]．中国农业气象，2023，44(11)：1057－1071.

[25] 周培诚，程塨，姚西文，等．高分辨率遥感影像解译中的机器学习范式[J]．遥感学报，2021，25(1)：182－197.

[26] 王敏钰，罗毅，张正阳，等．植被物候参数遥感提取与验证方法研究进展[J]．遥感学报，2022，26(3)：431－455.

[27] Feng X，Yu C，Chen Y，et al. Non-destructive determination of shikimic acid concentration in transgenic maize exhibiting glyphosate tolerance using chlorophyll fluorescence and hyperspectral imaging [J]. Frontiers in plant science，2018，9：468.

[28] 贺佳，王来刚，郭燕，等．基于无人机多光谱遥感的玉米LAI估算研究[J]．农业大数据学报，2021，3(4)：20－28.

[29] 韩文霆，彭星硕，张立元，等．基于多时相无人机遥感植被指数的夏玉米产量估算[J]．农业机械学报，2020，51(1)：148－155.
Han Wenting，Peng Xingshuo，Zhang Liyuan，et al. Summer maize yield estimation based on vegetation index derived from multi-temporal UAV remote sensing [J]. Transactions of the Chinese Society for Agricultural Machinery，2020，51(1)：148－155.

[30] 袁玉琦，陈瀚阅，张黎明，等．基于多变量与RF算法的耕地土壤有机碳空间预测研究——以福建亚热带复杂地貌区为例[J]．土壤学报，2021，58(4)：887－899.

[31] 潘根兴，丁元君，陈硕桐，等．从土壤腐殖质分组到分子有机质组学认识土壤有机质本质[J]．地球科学进展，2019，34(5)：451－470.

[32] Liu T，Xu T，Yu F，et al. A method combining ELM and PLSR (ELM－P) for estimating chlorophyll content in rice with feature bands extracted by an improved ant colony optimization algorithm [J]. Computers and Electronics in Agriculture，2021，186：106177.

[33] 张亚倩，骆社周，王成，等．联合无人机激光雷达和高光谱数据反演玉米叶面积指数[J]．遥感技术与应用，2022，37(5)：1097－1108.

[34] 谭先明，张佳伟，王仲林，等．基于PLS的不同水氮条件下带状套作玉米产量预测[J]．中国农业科学，2022，55(6)：1127－1138.

[35] 欧阳玲，毛德华，王宗明，等．基于GF－1与Landsat8 OLI影像的作物种植结构与产量分析[J]．农业工程学报，2017，33(11)：147－156，316.
Ouyang Ling，Mao Dehua，Wang Zongming，et al. Analysis crops planting structure and yield based on GF－1 and Landsat8 OLI images [J]. Transactions of the Chinese Society of Agricultural Engineering，2017，33(11)：147－156，316.

[36] 王宏轩，于珍珍，李海亮，等．基于GA－BP神经网络的鲜食玉米产量预测[J]．中国农机化学报，2024，45(6)：156－162.
Wang Hongxuan，Yu Zhenzhen，Li Hailiang，et al. Fresh corn yield prediction based on GA－BP neural network [J]. Journal of Chinese Agricultural Mechanization，2024，45(6)：156－162.

中国知网 https://www.cnki.net