中国农业大学学报 2022,27(5):134-145
Journal of China Agricultural University
http://zgnydxxb.ijournals.cn
DOI:10.11841/j.issn.1007-4333.2022.05.13

基于梯度提升算法的温室黄瓜株高生长模拟

翟子鹤[1] 陈小文[2] 高莉平[2] 张天柱[1*]
(1.中国农业大学 水利与土木工程学院,北京 100083;
2.北京中农富通园艺有限公司,北京 100083)

摘 要 为解决温室黄瓜株高生长模拟模型经验因子较多和实用性不强的问题,选取3个玻璃温室进行数据收集,以环境因子和时间因子为输入量,每日株高生长量为输出量,采用XGBoost模型建立黄瓜5个生育期的株高生长量模拟模型,并与LASSO模型进行对比分析,采用Pearson相关分析和XGBoost模型的特征重要性得分确定各生育期影响株高生长的重点因子。结果表明:1)XGBoost模型在黄瓜不同生育期的模拟性能均优于LASSO模型,苗期、伸蔓期、结果前期和结果末期表现出了较好的拟合效果,结果中期的拟合效果一般,苗期的模拟效果最好,决定系数为0.821;2)苗期、伸蔓期和结果期影响株高生长的重点因子分别是当前生育期生长天数、日平均湿度和日平均温度。本研究所建立的温室黄瓜株高生长模拟模型可为温室黄瓜生产环境调控优化提供决策支持。

关键词 黄瓜;温室;株高;XGBoost模型;模拟模型
中图分类号 S625.5 **文章编号** 1007-4333(2022)05-0134-12 **文献标志码** A

Simulation of greenhouse cucumber plant height growth based on gradient boosting algorithm

ZHAI Zihe[1], CHEN Xiaowen[2], GAO Liping[2], ZHANG Tianzhu[1*]
(1. College of Water Resources and Civil Engineering, China Agricultural University, Beijing 100083, China;
2. Beijing Zhongnong Futong Horticulture Co., Ltd., Beijing 100083, China)

Abstract In order to solve the problems of having many empirical elements and the weak generalization ability of plant height growth simulation model in greenhouse cucumber production, three glass greenhouses were selected to collect the data. Taking the environmental element and time element as input and the daily plant height growth as output, the simulation model of plant height growth in five growth stages of cucumber was established by XGBoost model and compared with LASSO model. Pearson correlation analysis and feature importance score of XGBoost model were used to determine the key elements affecting plant height growth in each growth stage. The results showed that: 1) The simulation performance of XGBoost fitting effect in full fruiting stage was general. The simulation effect in the seeding stage was the best, and the determination coefficient was 0.821. 2) The key elements affecting plant height growth in seedling stage, tendril elongation stage and fruiting stage were growth days in the current growth stage, daily average humidity and daily average temperature, respectively. In conclusion, the plant height growth simulation model of greenhouse cucumber established in this study can provide decision support for the regulation and optimization of greenhouse cucumber production environment.

Keywords cucumber; greenhouse; plant height; extreme gradient boosting model; simulation model

黄瓜的表型可以直观的反映出特定品种的生长状况,对黄瓜的精细管理、可视化建模和遗传改良具有重要的研究意义[1-5]。准确模拟黄瓜表型可以减少人工测量成本,跟踪监测作物长势,是黄瓜表型研

收稿日期:2021-08-09
基金项目:国家科技支撑计划项目(2011BAD12B03)
第一作者:翟子鹤,硕士研究生,E-mail:zhaizihehh@163.com
通讯作者:张天柱,教授,主要从事设施园艺环境工程研究,E-mail:zhangtianzhu@263.net

究中的重要方向。目前主要通过经验回归模型和图像学技术对黄瓜表型进行模拟[6-9]。但是经验回归模型在应用时需要确定较多的参数，应用性有限；基于图像学的模拟缺乏生物学基础[10]。株高是黄瓜的重要表型性状，与叶面积指数(LAI)有一定关系，同时株高也是黄瓜高产稳产的基础[11]。黄瓜株高与周围的小气候密不可分，有学者针对环境因子对黄瓜株高的影响进行了研究[12-14]，为种植者提供了管理依据，同时也为黄瓜株高生长模拟提供了研究基础。

温室黄瓜株高的模拟可以实时评估作物长势，了解环控效果，从而优化环控目标，许多学者对温室黄瓜株高模拟进行了研究。Kahlen 等[15]利用光量和叶面积作为黄瓜株茎节间变化的驱动因子，用回归分析的方法建立了黄瓜最终节间长度的模拟模型；李青林等[16]建立了黄瓜节间长度和节间直径的线性模型；李叶萌等[17]分别以有效积温、活动积温和辐射积作为株高的影响因子，建立了黄瓜株高的 Logistic 模型，结果显示利用辐热积指标模拟的精度较高。上述模型具有较强的机理性，但模型应用需要依赖较多的经验因子，实用性不强，而且实际生产中黄瓜株高的变化往往受多种因素的共同影响，难以用特定的数学函数关系来表达。机器学习是通过一定的算法从大量的历史数据中去学习规律，从而对新的样本去做预测或者分类[18]，不需要任何经验值，近年来已应用在部分作物株高的模拟中[19-21]。但是基于机器学习算法的温室黄瓜株高的模拟研究尚不多见。

XGBoost(Extreme gradient boosting)模型对于中小数据集具有较佳的预测表现，且具有算法可拓展性强、对异常值包容性强、并行速度快和加入正则项以防止过拟合等优点。本研究在 3 个连栋玻璃温室中进行温室环境数据和黄瓜株高生长数据的采集，采用 XGBoost 模型分别建立黄瓜 5 个生育期的株高生长量模拟模型，并与 LASSO 模型进行对比试验，同时进行黄瓜 5 个生育期各因子与株高生长量的相关性分析，并结合 XGBoost 模型的特征重要性确定各生育期影响株高生长的重点因子，为温室黄瓜生产环境调控优化提供决策支持，同时为进一步建立黄瓜的产量和品质模型奠定基础。

1　材料与方法

1.1　温室概况

试验地点位于河北省邢台市南和区的连栋温室群，该温室群由 1 个育苗温室，10 个生产温室，共 11 个连栋温室及中部连廊组成。温室顶部覆盖材料为钢化玻璃，温室立面为中空玻璃。温室立面底部的外墙材质为砖墙，高度为 1.0 m，厚度为 370 mm。单栋温室东西方向 19 跨，每跨长 9.6 m，共 182.4 m；南北方向共 13 个开间，每个开间 4.0 m，共 52.0 m。温室冬季的采暖方式为热水供暖，夏季采用湿帘-风机降温。栽培方式采用岩棉-椰糠复合栽培，行距为 1.6 m，株距为 25 cm。整枝方式为单秆整枝。灌溉方式为滴灌，在结果期之前和结果期分别使用不同的配方，每周根据基质的 EC/pH 进行浓度的动态调整，保证植株正常的营养供给。温室的环控由北京豪根道农业技术有限公司开发的气候环境控制系统(ISII)进行调控。本试验选用的温室为 6、7 和 8 号生产温室。

1.2　数据采集方法

单栋温室面积较大，温室不同方向的建筑结构和设备的不同会影响温室的保温性和光照分布的均匀性，进而造成温室不同区域内的环境存在较大差异，因此将温室划分为多个小区域，每个区域均布点测量。考虑到东西侧和南北侧的边际效应，东侧和西侧的传感器分别布置在距离东墙 19.2 m 和西墙 28.8 m 处，东西方向各传感器之间的距离为 67.2 m。南侧和北侧的传感器分别布置在距离南墙 8.0 m 和距离北墙 12.0 m 处，各传感器之间的距离为 16.0 m。每个温室设 9 个测点，3 个温室共 27 个测点。每个测点均采用普锐森社高精度温湿度传感器测量空气温湿度，采用普锐森社光照传感器测量光照强度，各传感器设备参数见表 1。温湿度传感器布置在椰糠条上方 1.5 m 处，并在其探头处套一个铝箔盒使其免受太阳辐射，光照传感器布置在植株上方，并且保证其不受遮挡。温湿度传感器和光照传感器的采集间隔均设置为 5 min。

黄瓜的品种为荷兰瑞克斯旺公司研制的‘冬之光’，该品种耐寒性好，抗病性强，适合于早春、早秋和秋冬温室栽培。黄瓜为三叶一心时定植，从定植后开始计算生育期。黄瓜自根部往上第 6 茎节处开始留果，由于黄瓜植株个体间存在差异，每个测点周围选择固定 3 棵试验植株进行株高测量，取其均值。由于黄瓜在温室中的长速较快，故设定测量频率为 1 天 1 次，测量时间为每天 8:00—9:00。株高前期用钢卷尺测量岩棉上表面至生长点的高度，待植株长到一定高度后用 30 cm 直尺直接测量其生长量。

表 1 传感器设备参数

Table 1 Sensor device parameters

传感器名称 Sensor name	型号 Type	量程 Range	精度 Accuracy	误差 Error
温度传感器 Temperature sensor	PR-3003-WS-5	−40～80 ℃	0.1 ℃	±0.1 ℃
传感器湿度 Humidity sensor	PR-3003-WS-5	0～100%	0.1%	±1.5%
光照传感器 Light sensor	PR-3002H-M-4G	0～20 万 lx	1 lx	±7.0%

试验共进行了 3 个播期的数据采集。7 号温室黄瓜的生长周期为 2020 年 8 月 28 日—2020 年 12 月 7 日,6 号温室黄瓜的生长周期为 2020 年 9 月 26 日—2021 年 1 月 7 日,8 号温室黄瓜的生长周期为 2020 年 10 月 29 日—2021 年 2 月 25 日。

1.3 数据预处理

生育期阶段划分:黄瓜在不同生育期的生长特性存在差异,为减小植株本体生长特性对模拟结果的影响,根据文献资料[22-23]和研究实际情况,将黄瓜的生长周期划分为 5 个阶段,分别建立株高生长量模拟模型,阶段划分依据如下:幼苗期:定植至第 4～5 片真叶展开;伸蔓期:第 4～5 片真叶展开至第一雌花完全开放;结果前期:第 1 雌花完全开放至根瓜采收;结果中期:根瓜采收至大量产瓜结束;结果末期:果实成熟缓慢至拉秧。将全部数据按照上述标准进行划分。

缺失值:光照传感器的数据传输方式为 GPRS,且需外接电源,由于园区内的断电或者传输信号的不稳定会造成数据缺失,故对于缺失值先用 3 次样条插值法进行插值,对插值后产生的个别负值进行剔除,并用线性插值法再次进行插值。

异常值:对光照异常数据采用均值法进行平滑修复,即:

$$x_k = \frac{x_{k-1} + x_{k+1}}{2} \tag{1}$$

式中:x_k 为异常数据,lx;x_{k-1} 和 x_{k+1} 为相邻有效数据,lx。

特征选择:由日常管理经验和文献资料可知,温室内每天温湿度的上下限、平均温度、平均湿度、平均光强、最大光强、生长天数和水肥条件对黄瓜的生长发育有不同程度的影响,但是水肥条件在温室中往往较易控制,且本试验中植物在黄瓜结果前和结果后分别采用固定的配方,故在本研究中水肥作为定量因子,不作为模型特征。同一生育期不同的生长阶段植株本体生长势不同,为减小由于植物当前生长势差异对株高生长量造成的影响,引入当前生育期生长天数这一特征,每一生长期的第 1 天测量标记为“1”,第 2 天测量标记为“2”,依次类推。最终选取的特征为:日平均光强、日最大光强、日平均温度、日最高温度、日最低温度、日平均湿度、日最大湿度、日最小湿度和当前生育期生长天数。

1.4 模型建立方法

梯度提升(Gradient boosting)算法是 Boosting 中的一大类算法,其基本原理是根据当前模型损失函数的负梯度信息来训练新加入的弱分类器,然后将训练好的弱分类器以累加的形式结合到现有模型中。采用决策树作为弱分类器的梯度提升算法被称为梯度提升树(Gradient boosting decison tree,GBDT)。XGBoost 是基于梯度提升树的一种集成算法,其基学习器为分类回归树,损失函数对误差部分进行了二阶泰勒展开,提升了精准度[24-25]。其目标函数分为 2 个部分,一部分是损失函数,一部分是正则化项(用于控制模型的复杂度,包括 L_1 正则化和 L_2 正则化)。目标函数表达式如下:

$$\mathrm{Obj} = \sum_{i}^{n} l(\hat{y}, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

式中:n 为样本数,个;$\hat{y}$ 为整个模型第 i 个样本的预测值,cm;y_i 为第 i 个样本的真实值,cm;K 代表全部树的数量,颗。

若使用 L_1 正则化,则正则化项展开式如下:

$$\sum_{k=1}^{K} \Omega(f_k) = \gamma T + \frac{1}{2}\alpha \mid \omega \mid \tag{3}$$

若使用 L_2 正则化,则正则化项展开式如下:

$$\sum_{k=1}^{K} \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \| \omega \|^2 \tag{4}$$

若同时使用 L_1 正则和 L_2 正则,则正则化项的展开式如下:

$$\sum_{k=1}^{K} \Omega(f_k) = \gamma T + \frac{1}{2}\alpha \mid \omega \mid + \frac{1}{2}\lambda \| \omega \|^2 \tag{5}$$

式中：γ 控制叶子数量，个；α 为 L_1 正则参数；λ 为 L_2 正则参数；ω 为决策树所有叶子节点值组成的向量；T 为叶子节点数，个。

XGBoost 模型中的各关键参数释义如下：N_estimators 是集成算法中弱评估器的数量，此参数值越大，模型的学习能力越强，但是模型过拟合的风险越大，一般以 300 以下为佳；Max_depth 为模型中树的最大深度，用于避免过拟合，此参数的值越大，代表模型越复杂，越容易过拟合，一般的取值范围为 3～10；Min_child_weight 控制叶子上所需的最小样本量，用于控制过拟合；Subsample 控制对于随机抽取的用于训练的数据的比例，典型值为 0.5～1.0；Learning_rate 为迭代速率，通过减小每一步的权重以提高模型的鲁棒性，典型值为 0.01～0.30。

LASSO 回归是在普通线性回归的目标函数后面加入了 L_1 范数惩罚项，能够同时实现变量选择和参数估计[26]。其目标函数如下：

$$\mathrm{Obj}=\frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}-y^{(i)})^2+\lambda\sum_{i=1}^{n}|\theta_i| \quad (6)$$

式中：m 为样本数，个；$h_\theta(x^{(i)})$ 为整个模型第 i 个样本的预测值，cm；$y^{(i)}$ 为整个模型第 i 个样本的实测值，cm；n 为参数数量，个；λ 为调整参数；θ 为回归系数。

本研究用 XGBoost 模型分别建立黄瓜 5 个生育期的株高生长量模拟模型，并与 LASSO 模型进行对比分析。模型的输入量为日平均光强、最大光强、日平均温度、日最高温度、日最低温度、日平均湿度、日最大湿度、日最小湿度和当前生育期生长天数，输出量为每日株高生长量。

根据每个生育期数据集的大小进行训练集和测试集的划分。其中，苗期、伸蔓期、结果前期和结果末期的数据集均按照 8∶2 随机划分为训练集和测试集，结果中期的数据集按照 9∶1 随机划分为训练集和测试集。

1.5　模型评估指标

采用决定系数 R^2，均方误差(MSE)，平均绝对误差(MAE)作为模型的评价分析指标。若模型的 R^2 越大，MSE 和 MAE 的值越小，则说明模型的拟合效果越好。

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (7)$$

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2 \quad (8)$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i| \quad (9)$$

式中：n 为样本数，个；$\hat{y}_i$ 为第 i 个样本的预测值，cm；$\bar{y}$ 为 n 个样本的平均值，cm；y_i 为第 i 个样本的真实值，cm。

1.6　影响株高生长重点因子的确定方法

相关性分析和模型特征重要性得分均可在一定程度上反映各因子对株高生长的影响程度。Pearson 相关系数法能在一定程度上反映因子间的相关性大小，但是由于其假设是线性相关，而实际上植物和环境因子之间并非完全的线性相关，所以不可仅仅依照 Pearson 相关系数法的结果确定影响株高生长的重点因子。XGBoost 模型的特征重要性得分是基于已经建立好的模型所得出的，特征重要性得分一方面可以反映某一特征对模型预测精度的影响，同时也可在一定程度上反映某一特征对因变量的影响，体现模型的可解释性，所以把这 2 种方法的结果放在一块可以更好的确定哪个因子对株高生长量的影响最大：即若某因子和株高的相关性很强，同时这一因子的特征重要性得分也很高，即可认为这一因子对株高生长的影响最大。

采用 SPSS 软件进行各因子间、各因子与株高的 Person 相关分析，得到相关系数。相关系数的取值范围为[−1,1]，相关系数的绝对值越大，代表相关性越强。基于已经建立好的各生育期 XGBoost 模型采用 Gain(信息增益的泛化概念，在 XGBoost 中指节点分裂时，该特征带来信息增益优化的平均值)获得特征重要性得分，特征重要性得分取值范围为[0,1]，得分越高，代表此特征对模型预测精度的影响越大。因子确定原则为：以相关性分析为主，取排名前 3 的相关性分析和特征重要性分析中的共有因子。

2　结果与分析

2.1　黄瓜株高生长量模拟模型的建立

机器学习的模型的预测效果一方面取决于数据本身的质量，另一方面取决于模型参数的调整。在调参过程中，首先进行 XGBoost 模型和 LASSO 各关键参数的范围设置，如表 2 所示。

由于 XGBoost 模型各关键参数的重要程度不

表 2 模型参数范围设置

Table 2 Range setting of model parameters

模型 Model	参数 Parameter	范围 Range
XGBoost	N_estimators	50～300
	Learning_rate	0.01～0.30
	Max_depth	3～7
	Min_child_weight	0～6
	Subsample	0.50～1.00
LASSO	Alpha	0～10

同，且部分关键参数互相之间的影响很大，故将关键参数分为 3 组，按调参顺序依次为：N_estimators 和 Learning_rate，Max_depth 和 Min_Chil_weight，Subsample。使用网格搜索和交叉验证的方法依次对上述 3 组的参数组合进行调整，每个参数在参数设置范围内选取 3～5 个候选值，调整过程中上组参数调到最优后在下一组中固定最优参数，然后进行调整，依次类推。LASSO 模型的关键参数 Alpha 为 L_1 正则化参数，用于控制模型的过拟合。经过多轮测试调参，最终确定的关键参数值如表 3 所示。

表 3 XGBoost 和 LASSO 的模型参数

Table 3 Model parameters of XGBoost and LASSO

生育期 Development stage	XGBoost					LASSO
	学习率 Learning_rate	最小样本和权重 Min_child_weight	最大树深 Max_depth	树的个数 N_estimators	随机采样比 Subsample	正则参数 Alpha
苗期 Seedling stage	0.01	6	6	250	0.74	0.10
伸蔓期 Tendril elongation stage	0.03	6	5	125	0.74	0.20
结果前期 Initial fruiting stage	0.03	6	3	200	0.85	1.92
结果中期 Full fruiting stage	0.03	3	5	200	0.74	0.05
结果末期 Last fruiting stage	0.03	4	3	210	0.94	0.01

由最终所得的模型得到在测试集上各生育期株高生长量的模拟曲线（图 1～5），由模拟曲线可以看出，在苗期、伸蔓期、结果前期和结果末期，XGBoost 模型的拟合效果较好。而在结果中期，XGBoost 模型的拟合效果一般，对于日平均株高生长量较多的少数点（＞8 cm），XGBoost 的模拟值明显偏小。LASSO 模型在 5 个生育期的拟合效果均较差，模拟性能均低于 XGBoost。LASSO 模型在结

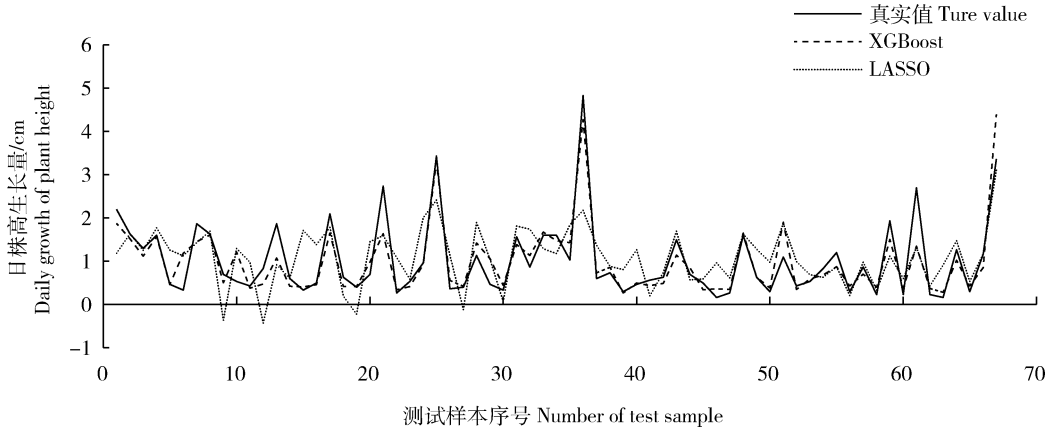


图 1 苗期株高生长量模拟曲线

Fig. 1 Simulation curve of plant height growth at seedling stage

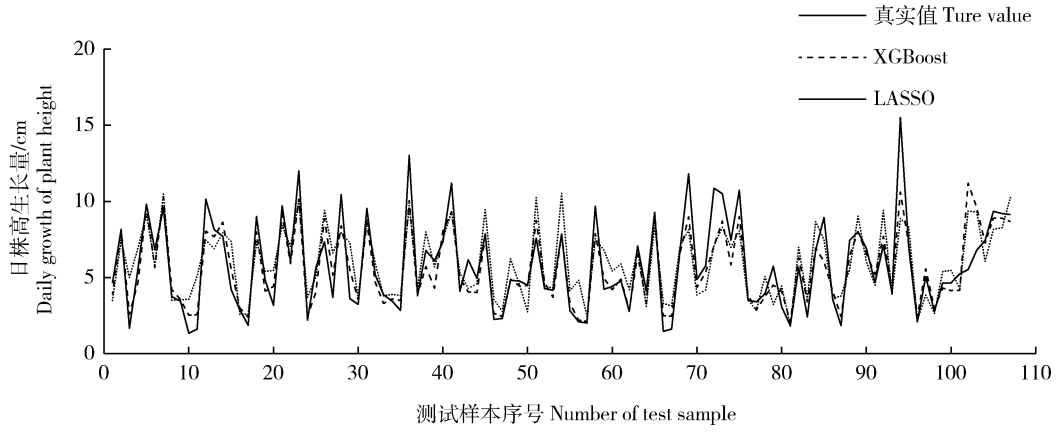


图 2　伸蔓期株高生长量模拟曲线

Fig. 2　Simulation curve of plant height growth at tendril elongation stage

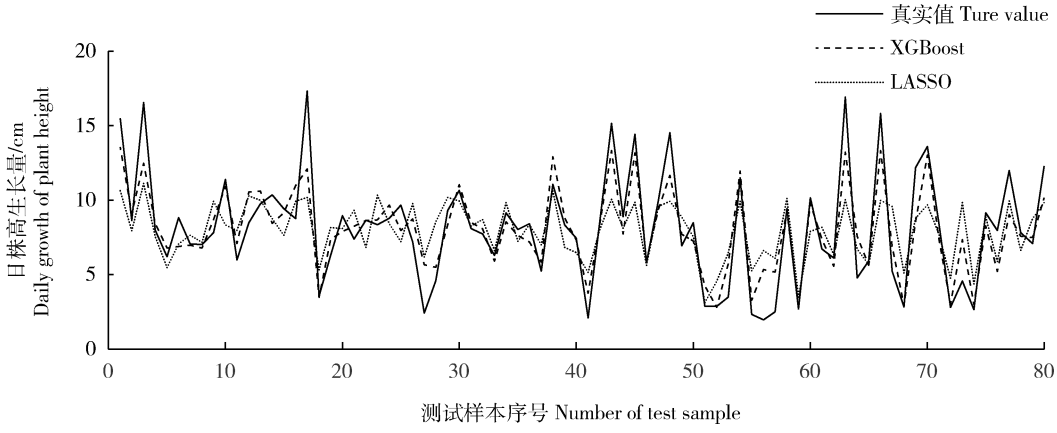


图 3　结果前期株高生长量模拟曲线

Fig. 3　Simulation curve of plant height growth at initial fruiting stage

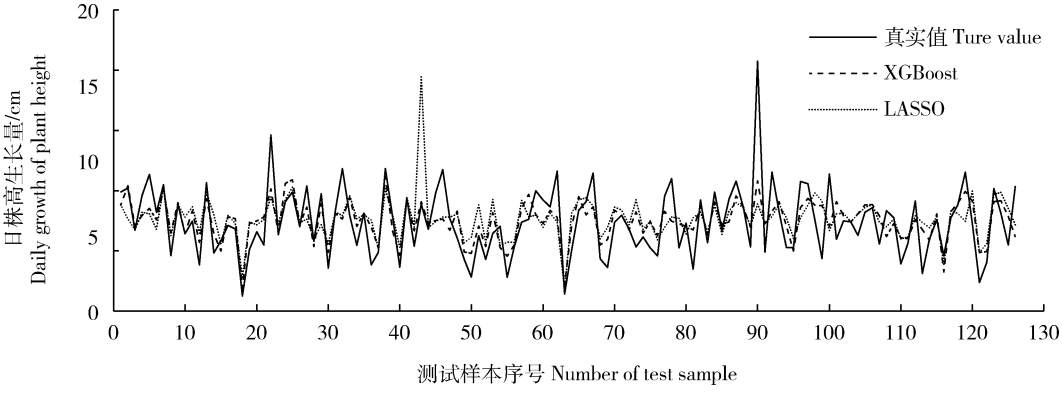


图 4　结果中期株高生长量模拟曲线

Fig. 4　Simulation curve of plant height growth at full fruiting stage

果中期的模拟效果在 5 个生育期中最差，对日平均生长量较多的点和日平均生长量较低的点均不能较好的拟合，个别点的模拟值和真实值的差距过大。

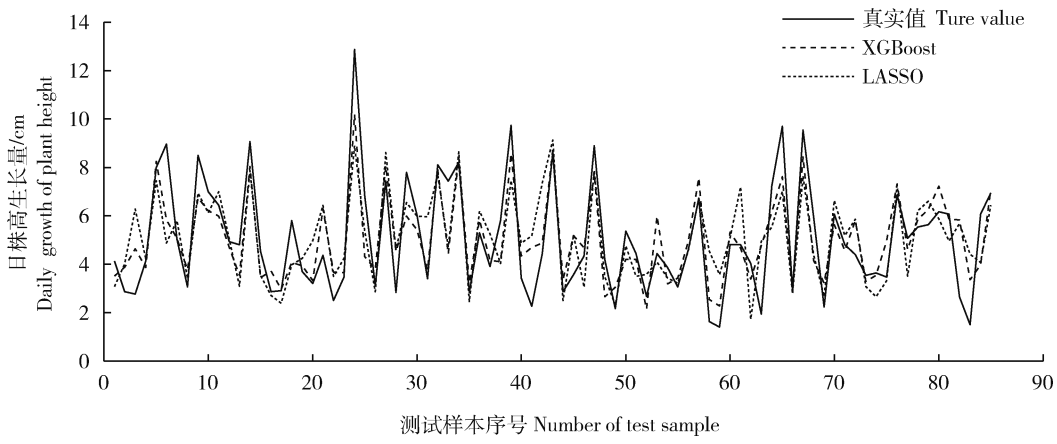


图 5 结果末期株高生长量模拟曲线

Fig. 5 Simulation curve of plant height growth at last fruiting stage

表 4 2 种模拟方法的结果评价

Table 4 Evaluation of two simulation methods

生育期 Development stage	模型 Model	R^2		MSE		MAE	
		训练 Train	测试 Test	训练 Train	测试 Test	训练 Train	测试 Test
苗期 Seedling stage	XGBoost	0.835	0.821	0.231	0.141	0.262	0.254
	LASSO	0.430	0.363	0.797	0.502	0.627	0.548
伸蔓期 Tendril elongation stage	XGBoost	0.893	0.805	1.112	1.717	0.754	0.892
	LASSO	0.655	0.643	3.584	3.135	1.455	1.387
结果前期 Initial fruiting stage	XGBoost	0.870	0.801	1.379	2.738	0.892	1.227
	LASSO	0.507	0.502	5.224	6.832	1.788	2.017
结果中期 Full fruiting stage	XGBoost	0.772	0.502	0.996	2.550	0.766	1.192
	LASSO	0.342	0.216	2.873	4.012	1.323	1.448
结果末期 Last fruiting stage	XGBoost	0.855	0.701	0.887	1.563	0.739	0.967
	LASSO	0.558	0.523	2.706	2.500	1.300	1.239

从模型的模拟效果来看，XGBoost 在 5 个生育期训练集和测试集的 R^2 均高于 LASSO，MSE 和 MAE 均低于 LASSO，表明 XGBoost 在整个生育期的模拟性能要优于 LASSO。XGBoost 在苗期、伸蔓期、结果前期和结果末期的测试集的 R^2 均大于 0.700，具有良好的模拟性能，其中苗期的模拟效果最好，测试集的 R^2 为 0.821。XGBoost 在结果中期的测试集 R^2 为 0.502，模拟性能一般。LASSO 模型在 5 个生育期的测试集的 R^2 均较低，模拟性能较差。

从模型的稳定性上来看，XGBoost 在苗期、伸蔓期、结果前期和结果末期的训练集和测试集的表现较为稳定，但在结果中期有一定波动，训练集和测试集的 R^2 的差距较大，表明模型存在轻度过拟合的现象。LASSO 在 5 个生育期的训练集和测试集的表现较为稳定。

2.2 影响株高生长的重点因子的确定

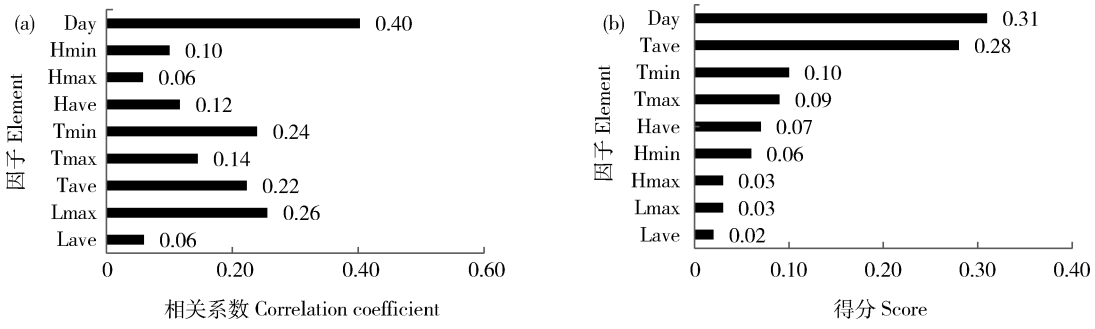
由于因子组合的共线性现象会影响相关性分析结果的可信度，故需要先确定各生育期因子组合是否存在共线性，以便后面更好的进行影响株高生长的重点因子的确定。考虑温室生产的实际情况，规定若因子间相关系数的绝对值＞0.85，即说明因子间存在共线性。采用 Person 相关数法进行分析，结果如表 5 所示。

表 5　各生育期的共线性因子

Table 5　Collinearity elements of every stage

生育期 Development stage	共线性因子 Collinearity element
苗期 Seedling stage	日平均温度和日最低温度、日平均湿度和日最大湿度、日平均湿度和日最小湿度
伸蔓期 Tendril elongation stage	无
结果前期 Initial fruiting stage	日平均湿度和日最大湿度
结果中期 Full fruiting stage	无
结果末期 Last fruiting stage	无

进行了各生育期因子间的共线性分析后即可进行影响株高生长的重点因子的确定(图 6～图 10)。苗期相关性分析和模型特征重要性得分见图 6。在苗期和株高生长量相关性排名前 3 的因子从高到低依次为:当前生育期生长天数、日最大光强和日最低温度,其中当前生育期生长天数的相关系数最大,为 0.40。当前生育期生长天数代表黄瓜在苗期不同阶段的长势情况。在苗期的初始,植株较小,整体长势较弱,随着生长天数的增加,植株叶片逐渐增加,植株自身的长势逐渐增强,株高的增量也逐渐增加。但是这一因子的值并非越大越好,若过大,表明植物可能存在徒长,影响开花结果,造成生殖生长和营养生长不平衡,这会影响植物的干物质积累,从而影响产量。因此在实际生产过程中,若随着生长天数的增加植物出现长速过快的情况,则需要及时采取措施来进行调整。同时当前生育期生长天数在模型的特征重要性得分中最高,这表明此特征对模型的预测结果影响最大,且由表 5 可知,当前生育期生长天数与其他因子间不存在共线性。故综合分析可得当前生育期生长天数是苗期影响株高生长的重点因子。



Lave 表示日平均光强,Lmax 表示日最大光强,Tave 表示日平均温度,Tmax 表示日最高温度,Tmin 表示日最低温度,Have 表示日平均湿度,Hmax 表示日最大湿度,Hmin 表示日最小湿度,Day 表示当前生育期生长天数。下同。

Lave represents the daily average light intensity, Lmax is the daily maximum light intensity, Tave is the daily average temperature, Tmax is the daily maximum temperature, Tmin is the daily minimum temperature, Have means daily average humidity, Hmax is the maximum daily humidity, Hmin represents the minimum daily humidity, Day represents the growth days of the current growth stage. The same below.

图 6　苗期株高与各因子相关性分析(a)和模型特征重要性得分(b)

Fig. 6　Correlation analysis between plant height and various elements (a) and feature importance score of model (b) at seedling stage

伸蔓期相关性分析和模型特征重要性得分见图 7。在伸蔓期和株高生长量相关性排名前 3 的因子从高到低依次为:日平均湿度、日平均温度和日最大湿度,其中日平均湿度的相关系数最大,为 0.63(日平均湿度的相关系数为 0.634 7,大于日平均温度的 0.634 5)。湿度可以影响作物蒸腾,而蒸腾作用是植物吸收水分和营养物质的动力,空气湿度过大会降低植株的蒸腾作用,导致营养物质的吸收和运输能力下降[27];空气湿度过低,会造成叶片边缘以及叶尖的坏死,进而影响植株生长。在伸蔓期,黄瓜的长势

较苗期显著增强，并由营养生长为主向生殖生长过渡，对水分和营养物质的需求加大，故此阶段日平均湿度对株高生长量的影响较大。同时日平均湿度在模型特征重要性得分中最高，且由表 5 可知伸蔓期日平均湿度与其他因子间不存在共线性。故综合分析可得日平均湿度是伸蔓期影响株高生长的重点因子。

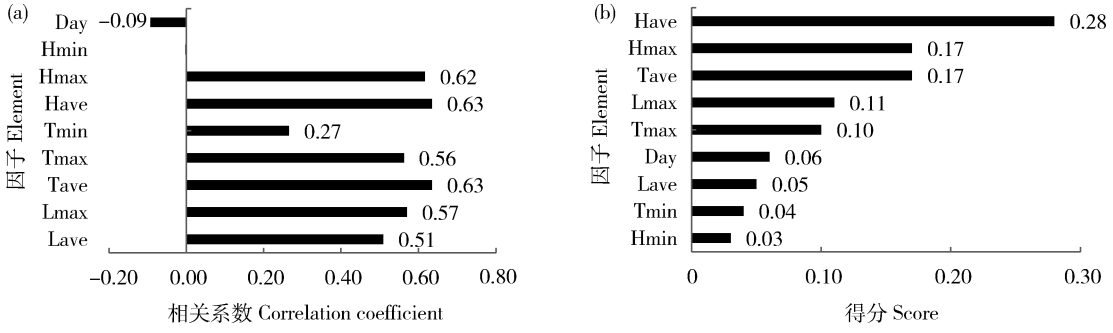


图 7 伸蔓期株高与各因子相关性分析(a)和模型特征重要性得分(b)

Fig. 7 Correlation analysis between plant height and various elements (a) and feature importance score of model(b) at tendril elongation stage

结果期相关性分析和模型特征重要性得分见图 8～10。在结果前期、中期和后期，和株高生长量相关性最强的因子均为日平均温度，相关系数分别为 0.72、0.55 和 0.49。温度对黄瓜的生长至关重要，结果期植株同时进行营养生长和生殖生长，且以生殖生长为主。适宜的温度可以让植株更加有效的进行光合作用，促进株高的生长，同时日平均温度这一特征在结果前期、中期和后期的特征重要性得分

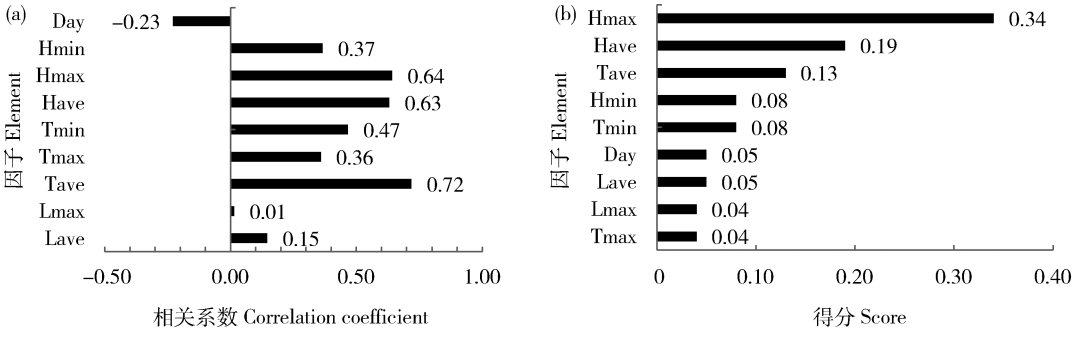


图 8 结果前期株高与各因子相关性分析(a)和模型特征重要性得分(b)

Fig. 8 Correlation analysis between plant height and various elements (a) and feature importance score of model(b) at initial fruiting stage

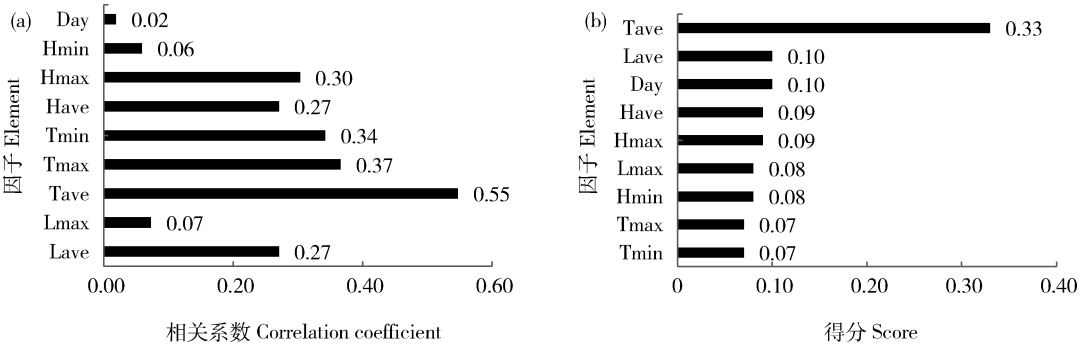


图 9 结果中期株高与各因子相关性分析(a)和模型特征重要性得分(b)

Fig. 9 Correlation analysis between plant height and various elements (a) and feature importance score of model(b) at full fruiting stage

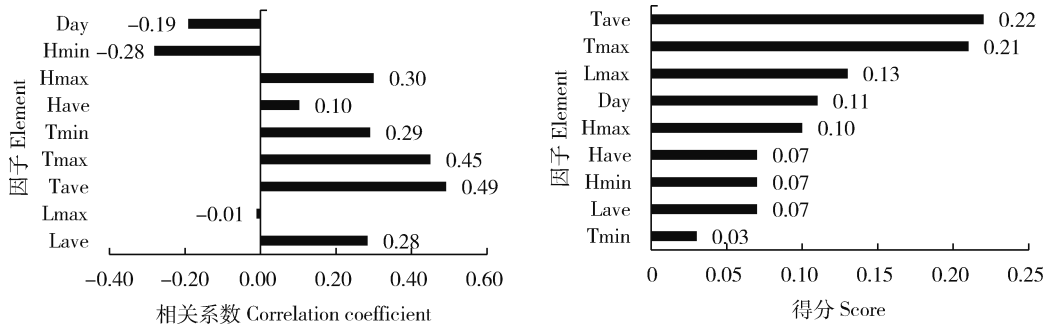


图 10　结果末期株高与各因子相关性分析(a)和模型特征重要性得分(b)

Fig. 10　Correlation analysis between plant height and various elements (a) and feature importance score of model(b) at last fruiting stage

排名依次为第 3、第 1 和第 1，且由表 5 可知，这 3 个生育期日平均温度与其他因子均不存在共线性。故综合分析可得日平均温度是结果期影响株高生长的重点因子。

3　讨论与结论

3.1　各生育期 XGBoost 模型的模拟性能

黄瓜的生长发育受温度、湿度和光照等多种环境因素的影响[17,28]，采用 1 种或 2 种环境因子对黄瓜生长进行模拟难免产生一定的模拟误差。本研究选取 3 个高产连栋玻璃温室，以温室内日平均光强、日最大光强、日平均温度、日最高温度、日最低温度、日平均湿度、日最大湿度、日最小湿度和当前生育期生长天数为输入量，黄瓜株高每日生长量为输出量，采用 XGBoost 模型建立了黄瓜不同生育期的株高生长量模拟模型，并与 LASSO 模型进行了对比分析。XGBoost 模型在黄瓜不同生育期的模拟性能均优于 LASSO 模型，苗期、伸蔓期、结果前期和结果末期表现出了较好的拟合效果，结果中期的拟合效果一般，本研究所建立的 XGBoost 模型的模拟性能整体上与李叶萌等[17]通过辐热积法建立的黄瓜株高模拟模型较为接近，但是 XGBoost 模型不依赖 3 基点温度等经验因子，实用性更强。XGBoost 模型在结果中期的模拟效果相对一般的原因可能是此阶段营养生长和生殖生长同时进行，株高的变化规律较其他时期更为复杂，模型对数据集的学习难度增加，导致模拟效果不如其他时期。从拟合曲线综合来看，XGBoost 模型对每个时期中绝大部分数据点都进行了较好的拟合，但是日株高生长量较多点 XGBoost 模型没有很好的拟合上，在结果中期的表现尤为明显，这可能是由于数据集本身的样本量不多，而增长量较多的样本在数据集中的占比很低，模型不易学习到这种变化规律。曾志雄等[29]利用 XGBoost 模型对猪舍温度预测的研究表明个别离群点和小规模数据集可能会降低模型的拟合效果，与本研究的结果类似。因此在后续研究中可增加样本数据量和优化特征工程，以进一步提高 XGBoost 模型的模拟性能。

3.2　各生育期影响株高生长的重点因子

在黄瓜的生长过程中，不同的环境因子对株高生长的影响程度在不断变化，确定黄瓜各个生育期影响株高生长的重点因子可以为温室黄瓜生产环境调控提供参考依据。本研究分别进行了黄瓜不同生育期株高生长量与各因子的相关分析，并结合各生育期已经建好的 XGBoost 模型的特征重要性得分来确定各生育期影响株高生长的重点因子。结果表明：苗期、伸蔓期和结果期影响株高生长的重点因子分别是当前生育期生长天数、日平均湿度和日平均温度。张帆洋等[30]研究发现在黄瓜全生育期的中后段，日平均温度和株高生长量有极显著的相关性，与本研究的研究结果一致。由于温室的环境因子和时间因子存在一定的耦合关系，这种耦合关系是温室内多种因素综合作用的结果，往往较为复杂，可能有线性和非线性等多种形式，本研究所得出的各生育期影响株高生长的重点因子在确定过程中只考虑了线性耦合形式，而对于非线性耦合对最终结果的影响还有待进一步研究。

3.3　XGBoost 模型的适用性

在中国现有的温室作物种植中，温湿度和光照是最为常见的环境监测因子，数据获取较为便利，同时由于 XGBoost 是基于数据驱动的模型，不依赖于经验因子，所以本研究所建立的温室黄瓜株高生长

模拟模型具有通用的潜力，也可为温室作物其他生长量的模拟提供研究思路。由于本研究是在可控温室内进行的，黄瓜的生长未受到营养胁迫和环境胁迫，若外部环境和水肥条件变化较大，模型的模拟性能则会受到一定影响，所以该模型在适应性和稳定性方面存在一定局限性。本研究所建立的温室黄瓜株高生长模拟模型主要针对特定的黄瓜品种和栽培条件，对不同黄瓜品种和栽培条件还有待进一步研究。

参考文献 References

[1] 方栋平，张富仓，李静，王海东，向友珍，张燕. 灌水量和滴灌施肥方式对温室黄瓜产量和品质的影响[J]. 应用生态学报，2015，26(6)：1735-1742
Fang D P，Zhang F C，Li J，Wang H D，Xiang Y Z，Zhang Y. Effects of irrigation amount and various fertigation methods on yield and quality of cucumber in greenhouse[J]. *Chinese Journal of Applied Ecology*，2015，26(6)：1735-1742 (in Chinese)

[2] Katrin K，Chen T W. Predicting plant performance under simultaneously changing environmental conditions—the interplay between temperature，light，and internode growth [J]. *Frontiers in Plant Science*，2015，6：1130

[3] Schmidt D，Kahlen K. Towards more realistic leaf shapes in functional-structural plant models[J]. *Symmetry*，2018，10(7)：278

[4] 唐卫东，刘欢，刘冬生，胡雪华，李萍萍，卢章平. 基于植株-环境交互的温室黄瓜虚拟生长模型研究[J]. 农业机械学报，2014，45(2)：262-268
Tang W D，Liu H，Liu D S，Hu X H，Li P P，Lu Z P. Virtual growth model of cucumber in greenhouse based on interaction between plant and environment[J]. *Transactions of the Chinese Society for Agricultural Machinery*，2014，45(2)：262-268 (in Chinese)

[5] 刘兴旺，翟许玲，张亚琦，尹帅，冯钟萱，任华中. 黄瓜果实形态建成的遗传及分子基础研究进展[J]. 园艺学报，2020，47(9)：1793-1809
Liu X W，Zhai X L，Zhang Y Q，Yin S，Feng Z X，Ren H Z. A review on genetic and molecular biology of fruit morphogenesis in cucumber[J]. *Acta Horticulturae Sinica*，2020，47(9)：1793-1809 (in Chinese)

[6] Kahlen K，Stüetzel H. Modelling photo-modulated internode elongation in growing glasshouse cucumber canopies[J]. *New Phytologist*，2011，190(3)：697-708

[7] 程陈，冯利平，薛庆禹，李春，宫志宏，董朝阳，伍露，王春雷，刘淑梅，李奕卓，黎贞发. 日光温室日光温室黄瓜生长发育模拟模型[J]. 应用生态学报，2019，30(10)：3491-3500
Cheng C，Feng L P，Xue Q Y，Li C，Gong Z H，Dong C Y，Wu L，Wang C L，Liu S M，Li Y Z，Li Z F. Simulation model for cucumber growth and development in sunlight greenhouse [J]. *Chinese Journal of Applied Ecology*，2019，30(10)：3491-3500 (in Chinese)

[8] 倪纪恒，陈学好，陈春宏，徐强，赵大球. 用辐热积法模拟温室黄瓜果实生长[J]. 农业工程学报，2009，25(5)：192-196
Ni J H，Chen X H，Chen C H，Xu Q，Zhao D Q. Simulation of cucumber fruit growth in greenhouse based on production of thermal effectiveness and photosynthesis active radiation[J]. *Transactions of the Chinese Society of Agricultural Engineering*，2009，25(5)：192-196 (in Chinese)

[9] 唐卫东，刘振文，刘冬生，龙满生. 基于形态重构的叶片性状特征可视化表达方法[J]. 农业机械学报，2019，50(8)：249-256，212
Tang W D，Liu Z W，Liu D S，Long M S. Visual expression method of leaf traits based on morphological reconstruction [J]. *Transactions of the Chinese Society for Agricultural Machinery*，2019，50(8)：249-256，212 (in Chinese)

[10] 周淑秋，郭新宇，雷蕾. 黄瓜生长可视化系统的设计与实现[J]. 计算机技术与发展，2007，17(1)：227-228，232
Zhou S Q，Guo X Y，Lei L. Design and realization of cucumber growing visualization system [J]. *Computer Technology and Development*，2007，17(1)：227-228，232 (in Chinese)

[11] 刘治开，牛亚晓，王毅，韩文霆. 基于无人机可见光遥感的冬小麦株高估算[J]. 麦类作物学报，2019，39(7)：859-866
Liu Z K，Niu Y X，Wang Y，Han W T. Estimation of plant height of winter wheat based on UAV visible image[J]. *Journal of Triticeae Crops*，2019，39(7)：859-866 (in Chinese)

[12] 于红，宫志宏，李春，李宁，吕玉环. 夜间低温对温室番茄及黄瓜苗期生长的影响[J]. 北方园艺，2017(3)：56-60
Yu H，Gong Z H，Li C，Li N，Lv Y H. Influence of low temperature at night on growth of tomato and cucumber at seedling stage in greenhouse[J]. *Northern Horticulture*，2017(3)：56-60 (in Chinese)

[13] 熊宇，刁家敏，薛晓萍，吕学梅，张继波. 持续寡照对冬季日光温室黄瓜生长及抗氧化酶活性的影响[J]. 中国农业气象，2017，38(9)：537-547
Xiong Y，Diao J M，Xue X P，Lv X M，Zhang J B. Effects of continuous overcast weather on cucumber growth and antioxidant enzyme activities in glasshouse[J]. *Chinese Journal of Agrometeorology*，2017，38(9)：537-547 (in Chinese)

[14] 刘金泉，严海欧，张清梅，候佳. CO_2 加富和短期昼间亚高温对温室嫁接黄瓜植株生长和光合作用的影响[J]. 北方园艺，2016(15)：50-54
Liu J Q，Yan H O，Zhang Q M，Hou J. Plant growth and photosynthesis of grafted cucumber in greenhouse under sub-high temperature and elevated CO_2 in short term daytime[J]. *Northern Horticulture*，2016(15)：50-54 (in Chinese)

[15] Kahlen K，Stüetzel H. Simplification of a light-based model

for estimating final internode length in greenhouse cucumber canopies[J]. *Annals of Botany*, 2011, 108(6): 1055-1063

[16] 李青林，毛罕平，李萍萍. 黄瓜地上部分形态-光温响应模拟模型[J]. 农业工程学报，2011，27(9)：122-127
Li Q L, Mao H P, Li P P. Simulation of cucumber organ above-ground with relation to light and temperature [J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2011, 27(9): 122-127 (in Chinese)

[17] 李叶萌，李冉，杨再强. 3 种光温指标在模拟设施黄瓜生长发育中的应用与比较[J]. 干旱气象，2013，31(3)：523-529
Li Y M, Li R, Yang Z Q. Application and comparison of three temperature and light indexes in simulating growth and development of cucumber in greenhouse[J]. *Journal of Arid Meteorology*, 2013, 31(3): 523-529 (in Chinese)

[18] 余凯，贾磊，陈雨强，徐伟. 深度学习的昨天、今天和明天[J]. 计算机研究与发展，2013，50(9)：1799-1804
Yu K, Jia L, Chen Y Q, Xu W. Deep learning: Yesterday, today, and tomorrow[J]. *Journal of Computer Research and Development*, 2013, 50(9): 1799-1804 (in Chinese)

[19] Osco L P, Junior J M, Ramos A P M, Furuya D E G, Santana D C, Teodoro L P R, Goncalves W N, Baio F H R, Pistori H, Junior C A S, Teodoro P E. Leaf nitrogen concentration and plant height prediction for maize using UAV-Based multispectral imagery and machine learning techniques [J]. *Remote Sensing*, 2020, 12(19): 3237

[20] Han L, Yang G J, Dai H Y, Xu B, Yang H, Feng H K, Li Z H, Yang X D. Modeling maize above-ground biomass based on machine learning approaches using UAV remote-sensing data [J]. *Plant Methods*, 2019, 15: 10

[21] 张瑜，汪小旦，孙国祥，李永博. 基于集合经验模态分解与 Elman 神经网络的线椒株高预测[J]. 农业工程学报，2015，31(18)：169-174
Zhang Y, Wang X C, Sun G X, Li Y B. Prediction of cayenne pepper plant height based on ensemble empirical mode decomposition and Elman neural network[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2015, 31(18): 169-174 (in Chinese)

[22] 李萍萍，周静，王纪章，付为国. 温室黄瓜生育期预测的正弦指数模型[J]. 江苏大学学报：自然科学版，2009，30(4)：325-329
Li P P, Zhou J, Wang J Z, Fu W G. Exponential sine equation for predicting cucumber growth process in greenhouses[J]. *Journal of Jiangsu University: Natural Science Edition*, 2009, 30(4): 325-329 (in Chinese)

[23] 曹元鑫，毕延刚，李娟起，高丽红，曲梅. 温室黄瓜发育期模拟模型的检验[J]. 中国农业大学学报，2014，19(3)：145-153
Cao Y X, Bi Y G, Li J Q, Gao L H, Qu M. Verification and evaluation of development stage model for greenhouse cucumber [J]. *Journal of China Agricultural University*, 2014, 19(3): 145-153 (in Chinese)

[24] 李占山，刘兆赓. 基于 XGBoost 的特征选择算法[J]. 通信学报，2019，40(10)：101-108
Li Z S, Liu Z G. Feature selection algorithm based on XGBoost[J]. *Journal on Communications*, 2019, 40(10): 101-108 (in Chinese)

[25] 程晓娜，孙志锋. 隐式反馈场景下的 LFM-XGB-LR 融合推荐算法[J]. 计算机工程与应用，2020，56(5)：85-92
Cheng X N, Sun Z F. LFM-XGB-LR hybrid recommendation algorithm in implicit feedback scenario [J]. *Computer Engineering and Applications*, 2020, 56(5): 85-92 (in Chinese)

[26] 刘建伟，崔立鹏，刘泽宇，罗雄麟. 正则化稀疏模型[J]. 计算机学报，2015，38(7)：1307-1325
Liu J W, Cui L P, Liu Z Y, Luo X L. Survey on the regularized sparse models[J]. *Chinese Journal of Computers*, 2015, 38(7): 1307-1325 (in Chinese)

[27] 李淑菊，丁圆圆，程智慧. 黄瓜耐冷耐湿性及其鉴定研究进展[J]. 中国蔬菜，2020(6)：23-30
Li S J, Ding Y Y, Cheng Z H. Research progress on cucumber cold and wet tolerance and its identification [J]. *China Vegetables*, 2020(6): 23-30 (in Chinese)

[28] 仝培江，程智慧，孟焕文. 黄瓜幼苗对高温高湿胁迫的生理响应[J]. 西北农林科技大学学报：自然科学版，2021，49(6)：85-93，103
Tong P J, Cheng Z H, Meng H W. Physiological response of cucumber seedlings to high temperature and humidity stress [J]. *Journal of Northwest A & F University: Natural Science Edition*, 2021, 49(6): 85-93, 103 (in Chinese)

[29] 曾志雄，罗毅智，余乔东，蔡任，吕恩利，夏晶晶. 基于时间序列和多元模型的集约化猪舍温度预测[J]. 华南农业大学学报，2021，42(3)：111-118
Zeng Z X, Luo Y Z, Yu Q D, Cai R, Lv E L, Xia J J. Temperature prediction of intensive pig house based on time series and multivariate models[J]. *Journal of South China Agricultural University*, 2021, 42(3): 111-118 (in Chinese)

[30] 张帆洋，王秀峰，黄雪，魏珉，杨凤娟，史庆华. 日光温室温光环境对黄瓜茎叶生长的影响[J]. 山东农业科学，2013，45(6)：44-47
Zhang F Y, Wang X F, Huang X, Wei M, Yang F J, Shi Q H. Effects of temperature and light on plant growth of cucumber in greenhouse[J]. *Shandong Agricultural Sciences*, 2013, 45(6): 44-47 (in Chinese)

责任编辑：董金波