文章编号：2095-2163(2024)02-0166-06 中图分类号：TP311.5 文献标志码：A

基于随机森林优化的神经网络算法在冬小麦产量预测中的应用研究

曾健铭，李 玥，魏霖静，赵 霞，周 慧
（甘肃农业大学 信息科学技术学院，兰州 730070）

摘 要：小麦产业涉及国家粮食安全和民生问题，通过对小麦产量进行科学准确的预测，对农业经济的发展、制定粮食进出口计划和确保国家粮食安全有重要意义。使用相关性分析遥感参数与产量之间的相关性，通过随机森林算法对特征变量进行重要性评价，剔除对目标相关性无关或影响较小的特征变量，最后，采用 BP 神经网络对产量进行预测。结果表明：归一化植被指数（Normalized Difference Vegetation Index，NDVI）在天水市整个冬小麦生育期内都与产量呈正相关关系；相对湿度、NDVI、最低温度、土壤湿度和辐照度为小麦产量预测的重要影响因子；与未进行特征变量筛选的情况相比，冬小麦产量预测的精准度显著提升，可以满足产量预测的精度要求，为相关的农业部门提供可靠的农情信息，为制定粮食政策与组织粮食生产提供参考依据。

关键词：随机森林；BP 神经网络；冬小麦；产量预测

Application of random forest optimized neural network algorithm in winter wheat yield prediction: a survey

ZENG Jianming, LI Yue, WEI Linjing, ZHAO Xia, ZHOU Hui
(College of Information Science and Technology, Gansu Agriculture University, Lanzhou 730070, China)

Abstract: The wheat industry is crucial for national food security and public welfare. Accurate and scientific prediction of wheat yield is significant for the development of agricultural economy, formulation of food import and export plans, and ensuring national food security. The method proposed in this paper uses correlation analysis to study the relationship between remote sensing parameters and yield. The random forest algorithm is employed to evaluate the importance of feature variables, eliminating those irrelevant or less impactful on the target correlation. Finally, the BP neural network is used for yield prediction. The results show that the Normalized Difference Vegetation Index (NDVI) has a positive correlation with yield throughout the entire winter wheat growth period in Tianshui City. Relative humidity, NDVI, minimum temperature, soil moisture, and irradiance are identified as important factors influencing wheat yield prediction. Compared to scenarios without feature variable selection, the accuracy of winter wheat yield prediction significantly improved, meeting the precision requirements for yield prediction. This provides reliable agricultural information for relevant agricultural departments and offers a reference for formulating grain policies and organizing grain production.

Key words: random forest; BP neural network; winter wheat; production forecast

0 引 言

中国作为人口大国和农业大国，粮食是人类生存之本，实现经济社会发展之基。小麦作为中国四大主粮之一，比重占粮食的三分之一，小麦产业是关系到国家粮食安全和民生的重要问题，通过对小麦产量进行科学准确的预测，对农业经济的发展、制定粮食进出口计划、确保国家粮食安全有重要意义。由于影响小麦产量的因素众多，不容易建立各影响因子与粮食产量的分析模型，对其精准的预测存在一定难度。

近年来，随着人工智能和遥感技术的快速发展，农业科技加速发展，为农业研究提供了新技术和新模式，推动了遥感对农业估产的发展[1]。目前遥感技术和遥感参数的作物估产方法主要有以下两种类型：一是作物模型，二是基于统计模型结合遥感参数的遥感估产方法。作物估产模型包括农业技术转移决策支持系统（DSSAT）[2]、农业生产系统模拟器

基金项目：国家自然科学基金（32060437，31360315）；甘肃农业大学青年导师基金项目（GAU-QDFC-2020-12）；甘肃省自然科学基金（18JR3RA165）。

作者简介：曾健铭（1996-），男，硕士研究生，主要研究方向：农业信息化研究。

通讯作者：李 玥（1979-），女，博士，副教授，主要研究方向：智慧农业、大数据分析与挖掘。Email：liyue@gsau.edu.cn

收稿日期：2023-02-24

(APSIM)[3]和世界粮食研究模型(WOFOST)[4]等等。以上模型需要输入的数据众多,如土壤数据、气象数据和施肥量等。虽然可以精确模拟作物生长过程,但是研究区域较小,遥感参数和作物模型结合的数据同化,可以实现大区域的产量估测,但是需要的数据量大及精细的数据,导致精准度不够高[5]。基于统计模型结合遥感参数的遥感估产方法包括线性和非线性模型,通常作物的产量表现是非线性的[6],因此非线性模型应用更加广泛,如随机森林[7-8]和神经网络[9-13]等。王来刚等[14]利用森林算法对特征变量进行了重要性分析和产量预测,得出增强型植被指数(Enhaneed Vegetation Index,EVI)、日光诱导叶绿素荧光(Sun-Induced Chlorophyll Fluorescence,SIF)和高程数据对小麦产量影响较大;刘峻明等[15]利用随机森林结合长时间序列气象数据,对冬小麦生育早期的产量预测取得良好的效果,得出平均温度、最低温度、负积温、最高温度在不同生育阶段对产量的影响程度;裴傲[16]将遥感数据和气象数据建立的神经网络预测玉米单产模型,证明了 NDVI、EVI、比值植被指数(Ratio Vegetation Index,RVI)和差值植被指数(Difference Vegetation Index,DVI)4 种植被数据以及气象数据,对产量影响的有效性和实用性;李海涛等[17]通过决策树筛选出最优的特征属性作为 BP 神经网络的输入参数,训练数据缩短,取得了良好的预测结果。

综上所述,本文针对输入特征变量筛选难和预测精度较低等问题,基于随机森林和 BP 神经网络,以天水市为研究区域,基于遥感参数和气象数据的结合与冬小麦实际总产量数据,使用随机森林重要性分析评估,对特征属性进行筛选,采用 BP 神经网络构建冬小麦产量预测模型,剔除对目标相关性无关或影响很小的特征属性,提升冬小麦的产量预测精度。

1　方法研究

1.1　随机森林算法

随机森林由多棵分类回归树(Classification and Regression Tree,CART)构建模型[18],其主要实现步骤如下:

(1)假设初始训练集为 N,通过自助法(Bootstrap)进行重采样,结合点随机分裂技术共同构建多棵决策树。随机采样过程中,将未被抽取的数据作为袋外数据(Out-of-Bag,OOB),使用抽取的 OOB 数据可估计局部误差和特征显著性评价;

(2)假设每一个样本有 M 个属性,决策树的每一个节点需要分裂时,随机从 M 个属性中抽取 m 个属性($m < M$),之后从 m 属性中采取某种策略(如:信息增益)选择一个最优的属性为该节点的分类变量;

(3)决策树分裂过程的每个节点都按照步骤 2 处理,直到不能再继续分裂(整个决策树形成过程不需要进行剪枝);

(4)由生成的多颗决策树组成的随机森林,将新的数据判别和分类,用不同的决策树投票来获取最终的分类结果。

1.2　特征变量重要性

原始数据集中往往有多个特征变量,从数据集中抽取一部分特征,使其降低特征维度提升算法性能,选择对结果影响较高的几个特征变量,以减少建模时特征变量数。随机森林模型不仅在预测问题上有着广泛的应用,还可以对特征变量进行重要性分析。本研究通过随机森林分析 OOB 误差评价特征变量,对高维数据样本进行筛选,从而得到各特征变量的重要性,选择重要性较高的几个作为 BP 神经网络的输入变量。计算特征变量重要性的具体步骤如下:

(1)使用对应的 OOB 数据,计算每颗决策树的袋外数据误差(记作 errOOB1)。这样每棵决策树都得到一个 errOOB1,T 棵决策树就有 T 个 errOOB1。

(2)遍历所有特征,考察特征的重要性。随机对袋外数据样本特征遍历,并随机更改特征变量值(该操作加入噪声干扰),然后重新计算袋外数据误差(记为 errOOB2)。

(3)当随机森林中有 T_{tree} 颗树时,特征变量的重要性用公式 $\sum (e_{\text{OOB2}} - e_{\text{OOB1}}) / T_{\text{tree}}$ 表示。若随机给某个特征变量加入噪声干扰,则袋外准确率将大幅降低,表明该特征变量对于模型预测结果影响很大,同时说明该特征变量重要程度较高。

1.3　BP 神经网络

BP 神经网络(Back-Propagation Network)1986 年由 Rumelhart 和 McClelland 为首的科学家小组提出,是目前应用于产量预测最广泛的神经网络模型之一。BP 神经网络按误差反向传播算法训练,主要由输入层、输出层以及一个或多个隐含层组成,其网络结构如图 1 所示。BP 神经网络的输入为 x_i;w_{ij} 为输入层与隐含层的权值;Φ 为隐含层激活函数;w_{id} 为隐含层与输出层之间的权值;输出层激活函数为 Ψ;θ_i、θ_k 分别为隐含层与输出层的阈值;θ_k 为神经网络的输出[19]。

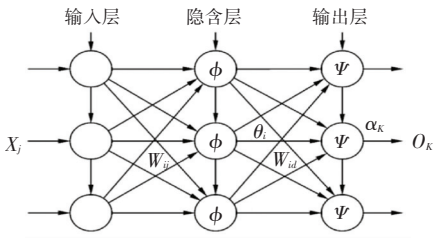

图 1 3 层 BP 神经网络结构图

Fig. 1 Structure diagram of 3-layer BP neural network

1.4 预测模型构建

在相关数据输入 BP 神经网络模型之前,需将遥感参数和气象数据,通过随机森林重要性评估方法,剔除多余的特征变量,选取最优的特征变量集合作为 BP 神经网络输入节点构建神经网络,并在特征变量属性和冬小麦产量之间建模,如图 2 所示。

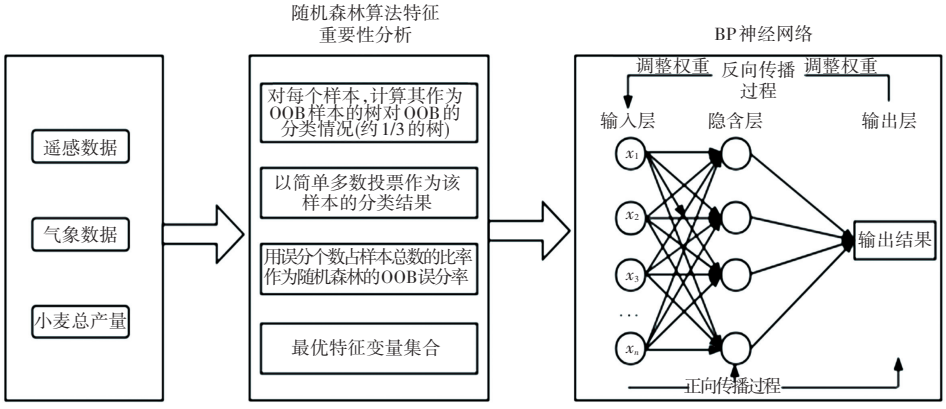

图 2 产量预测流程

Fig. 2 Yield prediction process

经过对特征变量数据训练,获取预测网络,通过测试和调整,对冬小麦产量进行预测。主要步骤如下:

(1)为了使数据的量纲保持一致,将数据统一到[0-1]之间,归一化公式为

$$x_i = \frac{x - x_{\max}}{x_{\max} - x_{\min}} \tag{1}$$

其中,x 为初始数据;$x_{\min}$、$x_{\max}$ 分别为初始数据的最小值和最大值;x_i 为归一化处理后的数据。

(2)通过对特征变量进行随机森林的 OOB 重要性分析,结合模型情况确定网络最佳特征变量集合。

(3)将最佳特征变量集合作为 BP 神经网络模型输入,确定网络结构和隐含层数进行训练。

(4)对网络进行测试验证,查验训练效果和预测精准度是否达到预期,如达到最大迭代次数和精准度,则停止网络训练并获取输出结果。

1.5 参数设计

在 BP 神经网络的输入层,输入由随机森林的重要性评估后的相对湿度、NDVI、最低温度、土壤湿度和辐照度等 5 个影响产量的特征变量值,将小麦产量作为 BP 神经网络模型网络输出。本文神经网络输入层到隐含层采用 Relu 函数,输出层采用 linear 函数,学习速率为 0.000 1,训练次数为 2 000 次。

在 BP 神经网络中,输入层和输出层的节点数都是确定的,而隐含层节点数是根据经验公式确定[20],计算公式为

$$h = \sqrt{m + n} + a \tag{2}$$

式中:h 为隐含层节点的神经元数,m 和 n 分别是输入层和输出层节点的神经元数,a 为 1 - 10 之间的调节常数。根据公式隐含层确定在 7 - 16 之间,依据不同隐层节点数训练结果比较,本文选择隐含层的节点数为 16。

2 验证分析

2.1 数据来源

试验所需数据包括 2000-2021 年天水市各县冬小麦生长期的遥感、气象和小麦产量数据。

2.1.1 遥感数据

遥感参数采用归一化植被指数(Normalized Difference Vegetation Index,NDVI),其是反映作物长势和营养信息的重要参数,与作物的产量有很好的相关性,常被用于产量预测的特征变量[21-22]。本文选取的天水市各县植被指数均来自美国国家航空航天局(NASA),MOD13Q1 产品空间分辨率是250 m,时间分辨率是 16 d,并按天水市耕地进行掩膜处理,

将 MODIS 图像在 Arcgis 软件中进行波段运算，得到天水市各县 2000-2021 年（每年 10 月-次年 5 月份）NDVI 植被指数的分布情况，采用最大值合成法得到每个月的最大植被指数数据。

2.1.2　气象数据

气象数据来自 NASA Power 气象数据库获取的 2000-2021 年天水市气象要素。气象要素来自天水市麦积区、甘谷、秦安、秦州、清水、武山和张家川 7 个区县的数据，其中包括辐照度、最高气温、最低温度、土壤湿度、相对湿度、平均气温和降雨量等 7 个要素。

2.1.3　小麦产量数据

天水市各县 2000-2021 年的小麦产量数据来源于《甘肃发展年鉴》。

2.2　评价指标

实验中采用平均绝对百分误差（Mean Absolute Percentage Error，*MAPE*）、均方根误差（Root Mean Squared Error，*RMSE*）和平均绝对误差（Mean Absolute Error，*MAE*）作为评价指标，对预测模型的性能进行比较。计算公式如下：

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - p_i}{y_i}\right| \times 100\% \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - y_i| \quad (5)$$

式中：p_i 是小麦产量的预测值，y_i 是实际值。*MAPE*、*RMSE* 和 *MAE* 的值越小，说明预测值与实际值偏差越小，预测性能越好，反之说明预测性能越差。

2.3　植被指数与产量的相关性分析

为研究植被指数和冬小麦产量之间的关系，从时间上对归一化植被指数（Normalized Difference Vegetation Index，NDVI）与产量之间进行相关性分析。遥感参数（NDVI）与小麦产量在每个月份之间的相关性如图 3 所示。在整个冬小麦生育期，NDVI 与产量都呈正相关，在冬小麦生长关键期 2-5 月份，相关系数均达到 0.4 左右；2-4 月份达到了最高峰，该期间属于冬小麦返青-孕穗期，此时小麦进入了旺盛的生长期，营养生长与生植生长并进的重要时期。在此期间，生长所需的水分和养分最多，叶面积及茎穗快速增长，直接决定了穗数和粒数的关键阶段，也是影响小麦产量高低最关键时期。之后，因为小麦冠层叶片衰老和籽粒灌浆，NDVI 与小麦产量之间的相关性降低。

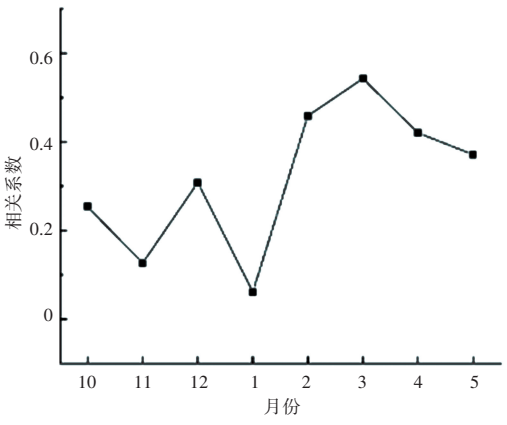

图 3　NDVI 与冬小麦产量相关性

Fig. 3　The correlation between NDVI and winter wheat output

2.4　特征变量重要性分析

特征选择不仅可以防止模型过拟合、减少模型的泛化误差，还可以减少硬件资源的损失、模型的开发成本和训练时间。有些特征变量对目标相关性低或者无关，输入的特征变量属性过多将导致网络收敛速度降低，从而增加过拟合的几率。因此，对神经网络训练前将特征变量进行筛减，选取重要性较高的 5 个特征变量作为 BP 神经网络的特征集。将 NDVI、辐照度、相对湿度、土壤湿度、降水量、最高温度、最低温度和平均气温等特征变量，采用随机森林的袋外 OOB 进行重要性分析，特征变量重要性指标由大到小排序（见图 4）。分析表明，相对湿度、NDVI、最低温度、土壤湿度和辐照度对小麦产量的重要性大于其它因素，重要性指标平均值都超过了 0.13，说明这些特征变量是影响小麦产量的重要环境因子；而降雨量、最高温度和平均温度的重要性相对较低，对小麦产量的影响较小，因此可将这些特征变量剔除。

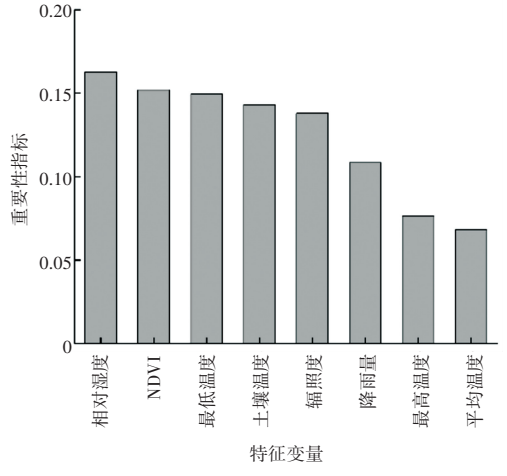

图 4　小麦特征变量重要性统计图

Fig. 4　Statistics of the importance of variables in wheat characteristic

2.5 模型预测结果分析

本文选取 2000-2021 年的数据作为实验样本数据。为了验证该模型的预测精准度,将样本数据分为训练样本和测试样本两部分。其中 2000-2018 年的数据作为训练集数据用于模型训练,利用训练好的模型对 2019-2021 年的产量进行预测,将结果与年鉴中的实际小麦产量数据进行对比与分析,结果见表 1。

表 1 模型预测结果

Table 1 Prediction results

年份	天水市各县	真实值(kg/hm²)	预测数据(kg/hm²)	绝对误差(kg/hm²)	相对误差(%)
2019	甘谷	3 263.03	3 337.11	74.08	2.27
2019	麦积	3 305.92	3 146.03	159.89	4.84
2019	秦安	2 976.59	3 143.96	167.37	5.62
2019	秦州	3 334.80	3 133.36	201.44	6.04
2019	清水	3 099.64	3 088.17	11.47	0.37
2019	武山	3 233.83	2 826.04	**407.79**	12.61
2019	张家川	2 798.94	3 156.68	357.74	**12.78**
2020	甘谷	3 364.75	3 432.13	67.38	2.00
2020	麦积	3 461.16	3 462.13	**0.97**	**0.03**
2020	秦安	3 109.37	3 359.63	250.26	8.05
2020	秦州	3 458.59	3 602.49	143.90	4.16
2020	清水	3 204.23	3 445.24	241.01	7.52
2020	武山	3 361.89	3 327.79	34.10	1.01
2020	张家川	2 914.36	3 248.58	334.22	11.47
2021	甘谷	3 520.62	3 466.77	53.85	1.53
2021	麦积	3 650.76	3 390.41	260.35	7.13
2021	秦安	3 270.65	3 341.11	70.46	2.15
2021	秦州	3 620.77	3 361.94	258.83	7.15
2021	清水	3 355.98	3 425.15	69.17	2.06
2021	武山	3 453.98	3 097.18	356.80	10.33
2021	张家川	3 052.02	3 256.70	204.68	6.71

结果表明,小麦的估测数据与年鉴中的实际小麦产量数据之间的绝对误差最高值是407.79 kg/hm²,绝对值的最低值是 0.97 kg/hm²,平均绝对误差值是 177.42 kg/hm²;相对误差值最低的是 0.03%,最高值是 12.78%,平均相对误差值是 5.52%,说明模型的预测结果满足了对小麦产量预测的要求,能较好的对研究区的小麦产量进行预测。

表 2 不同模型的精度评价

Table 2 Accuracy evaluation of different models

模型	*MAPE*/ %	*RMSE*/ (kg · hm^{-2})	*MAE*/ (kg · hm^{-2})
BP 神经网络	11.31	490.28	400.61
本文模型	6.91	214.86	177.41

根据表 2 对比结果表明,本文模型的 *MAPE* 为 6.91%、*RMSE* 为 214.86 kg/hm²、*MAE* 为177.41 kg/hm²,而 BP 神经网络的 *MAPE* 为 11.31%、*RMSE* 为 490.28 kg/hm²、*MAE* 为 400.61 kg/hm²。相比之下,对于冬小麦产量预测精准度有明显提升。

3 结束语

针对小麦产量预测问题,分析遥感参数与产量之间的相关性,构建了基于随机森林和 BP 神经网络组合的小麦产量预测模型。该模型基于遥感数据、气象数据和产量统计数据,所需的数据简单易得,并且能够有较高的预测精准度,为相关的农业部门提供可靠的农情信息,为制定粮食政策与组织粮食生产提供参考依据。结合实际数据,得出以下结论:

NDVI 是小麦产量预测的重要因子,与小麦产量呈高度相关性,特别是在冬小麦生长关键期 2-5 月份达到了最高,相关系数均达到 0.4 左右。说明 NDVI 是评估小麦生长和产量的重要指标。在 8 类

特征变量中，通过本文模型的筛选可以得出，相对湿度、NDVI、最低温度、土壤湿度和辐照度这几类特征变量对冬小麦的产量影响较大，其次是降雨量、最高温度和平均温度对小麦产量影响的重要程度相当，重要性相对较低。选择遥感数据和气象数据结合作为特征变量，对历史数据进行训练，进而预测 2019-2021 年的产量，与未进行特征变量筛选的情况相比，冬小麦产量预测的精准度显著提升，可以满足产量预测的精度要求。

本文的冬小麦产量预测模型输入变量主要考虑了植被的生长状态和气象条件，然而在实际生产过程中还存在诸多的影响因素（如品种、播种量和病虫害等因素），对产量也有很大的影响。未来建模方法当中，在遥感参数方面可以使用增强植被指数 EVI[23]，因为引入了蓝光波段对大气气溶胶的散射和土壤背景进行了矫正，减弱了来自大气和土壤的噪声影响，可以更好地反应被测地区的植被指数的情况，从而提高模型的精准度。

参考文献

[1] 史舟，梁宗正，杨媛媛，等. 农业遥感研究现状与展望[J]. 农业机械学报，2015，46(2)：247-260.

[2] 杨勤，许吟隆，林而达，等. 应用 DSSAT 模型预测宁夏春小麦产量演变趋势[J]. 干旱地区农业研究，2009，27(2)：41-48.

[3] ASSENG S，KEATING B A，FILLERY I R P，et al. Performance of the APSIM-wheat model in Western Australia[J]. Field Crops Research，1998，57(2)：163-179.

[4] VANDIEPEN C A，WOLF J，VAN KEULEN H，et al. WOFOST：A simulation model of crop production[J]. Soil Use and Management，1989，5(1)：16-24.

[5] 王静，李新. 基于作物生长模型和多源数据的融合技术研究进展[J]. 遥感技术与应用，2015，30(2)：209-219.

[6] 朱再春，陈联裙，张锦水，等. 基于信息扩散和关键期遥感数据的冬小麦估产模型[J]. 农业工程学报，2011，27(2)：187-193，398.

[7] 林滢，邵怀勇. 基于随机森林算法的河南省冬小麦产量预测最佳时间窗和影响因子研究[J]. 麦类作物学报，2020，40(7)：874-880.

[8] 王鹏新，齐璇，李俐，等. 基于随机森林回归的玉米单产估测[J]. 农业机械学报，2019，50(7)：237-245.

[9] 江东，王建华，杨小唤，等. 应用神经网络建立冬小麦产量预测模型[J]. 农业系统科学与综合研究，1999(2)：95-97.

[10] 周亮，慕号伟，马海姣，等. 基于卷积神经网络的中国北方冬小麦遥感估产[J]. 农业工程学报，2019，35(15)：119-128.

[11] JI B，SUN Y，YANG S，et al. Artificial neural networks for rice yield prediction in mountainous regions[J]. The Journal of Agricultural Science，2007，145(3)：249-261.

[12] Monisha Kaul，Robert L Hill，Charles Walthall. Artificial neural networks for corn and soybean yield prediction[J]. Agricultural Systems，2004，85(1)：1-18.

[13] WANG A X，TRAN C，DESAI N，et al. Deep transfer learning for crop yield prediction with remote sensing data[C]// Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies. 2018：1-5.

[14] 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.

[15] 刘峻明，和晓彤，王鹏新，等. 长时间序列气象数据结合随机森林法早期预测冬小麦产量[J]. 农业工程学报，2019，35(6)：158-166.

[16] 裴傲. 基于神经网络的玉米遥感估产模型研究[D]. 长春：吉林农业大学，2021.

[17] 李海涛，刘泰麟，邵泽东，等. 基于决策树与二分分割算法的 BP 神经网络在赤潮等级预测中的应用研究[J]. 海洋科学，2019，43(9)：34-40.

[18] BREIMAN L. Random forests[J]. Machine Learning，2001，45：5-32.

[19] 马湧，王晓鹏，马莎莎. 基于 Keras 深度学习框架下 BP 神经网络的热轧带钢力学性能预测[J]. 冶金自动化，2019，43(2)：6-10.

[20] 沈花玉，王兆霞，高成耀，等. BP 神经网络隐含层单元数的确定[J]. 天津理工大学学报，2008，24(5)：13-15.

[21] JOHNSON M D，HSIEH W W，CANNON A J，et al. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods[J]. Agricultural and Forest Meteorology，2016，218-219：74-84.

[22] BUKHOVETS A G，SEMIN E A，KUCHERENKO M V，et al. Forecasting the winter wheat yield based on the vegetation index NDVI dynamic model[J]. IOP Conference Series：Earth and Environmental Science，2021，848(1)：012191.

[23] 唐俊，赵成萍，周新志，等. 基于 EVI-RBF 的玉米长势监测及产量预测[J]. 江苏农业学报，2020，36(3)：577-583.