结合 GWRFR 和作物物候信息的玉米产量早期预测

裴 杰[1,2]，谭绍锋[1]，郭 韩[1]，刘一博[1]，方华军[3,4※]

（1. 中山大学测绘科学与技术学院，珠海 519082；2. 自然资源部华南热带亚热带自然资源监测重点实验室，珠海 519082；
3. 中国科学院地理科学与资源研究所生态系统观测与模拟重点实验室，北京 100101；
4. 中科吉安生态环境研究院，吉安 343000）

摘 要：及时并准确地估计作物产量，对保障粮食安全、维护世界粮食供应稳定具有重要意义。此前，已有许多研究者使用机器学习方法对作物产量预估进行研究。然而，结合作物的空间分布、使用局部模型进行分析的研究较少；且诸多研究均以年份为时间尺度进行建模，未能精细到作物生长的各个阶段，无法实现作物产量的早期预测。针对以上问题，该研究结合多源遥感数据，利用随机森林（random forest，RF）以及地理加权随机森林（geographically weighted random forest regression，GWRFR）模型对美国县级玉米产量进行建模，探讨全局与局部模型在玉米产量预测方面的性能；并通过将 GWRFR 模型应用于玉米的各个物候期，获取了玉米产量的最佳提前预测时间。结果表明，GWRFR 局部模型的精度（R^2=0.87，RMSE=864.21 kg/hm^2）高于传统的 RF 全局模型（R^2=0.83，RMSE=994.75 kg/hm^2），并且能够较好地克服空间数据的非平稳性，即使在全局模型中加入经纬度作为变量，RF 模型的预测效果（R^2=0.85，RMSE=890.88 kg/hm^2）仍然低于 GWRFR 模型。对于玉米产量的预测可以提前至收获前 2～3 个月，即在乳熟期前后就能得到比较准确的预测结果（R^2=0.90，RMSE=748.39 kg/hm^2）。该研究结果可为大尺度作物产量预估提供一种新的思路，对区域或全球其他作物的产量预测也具有一定的指导意义。

关键词：产量；预测；遥感；机器学习；作物物候
doi：10.11975/j.issn.1002-6819.202308028
中图分类号：S127 **文献标志码：**A **文章编号：**1002-6819(2024)-01-0161-09

裴杰，谭绍锋，郭韩，等. 结合 GWRFR 和作物物候信息的玉米产量早期预测[J]. 农业工程学报，2024，40(1)：161-169. doi: 10.11975/j.issn.1002-6819.202308028 http://www.tcsae.org
PEI Jie, TAN Shaofeng, GUO Han, et al. Early prediction of maize yield by integrating GWRFR and crop phenological information[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(1): 161-169. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.202308028 http://www.tcsae.org

0 引 言

粮食安全是维系社会稳定和保障社会发展的重要支撑。然而，由于全球人口增长、气候变化、耕地占用、土地退化等众多因素的影响，粮食生产和供求形势面临极大风险，保障粮食安全已经成为国际社会的共识[1]。作物产量预测在农业规划管理中起着关键的作用，其对地区乃至全球范围内的农业生产和粮食安全均具有重要意义[2-3]。及时、准确预测粮食产量可以帮助政府对粮食生产中的方案和策略进行调整，有利于提高粮食生产的效率，提升现代化农业的发展水平[4]。玉米作为世界三大粮食作物之一，主要用于食品、饲料以及工业加工等领域，具有广泛的应用价值[5]。因此，及时并准确地对玉米产量进行大规模预测，能为政府对粮食生产和经济政策采取宏观调控决策提供基础信息，有助于保障粮食安全、维护世界粮食供应的稳定[6]。

当前主流的作物产量预测方法主要包括两类：基于作物生长过程的物理模型和基于机器学习的统计模型[7]。物理模型以作物的生理特性为基础，通过模拟潜在的作物和环境进程（如作物生长、养分循环、水平衡）来估计产量[8]，其中 DSSAT（decision support system for agrotechnology transfer）使用最为广泛[9]。但是，该模型在进行长期模拟之前，需要大量的实地观测数据对模型进行校准，这限制了该方法在大区域范围产量预测中的适用性[10]。机器学习作为近几年来兴起的一种新方法，能够从输入数据中直接学习相关信息，并通过建立产量驱动因素与历史产量记录之间的经验关系来进行产量估计，具有不依赖于作物参数的优势，已经广泛用于作物产量的预测[11]。例如，严海军等[12]利用支持向量回归（support vector regression，SVR）进行苜蓿产量预测。刘峻明等[13]基于随机森林算法对冬小麦产量进行回归预测，预测误差在 10% 以内。孙少杰等[14]将遥感数据与卷积神经网络（convolutional neural networks，CNN）和反向传播神经网络（back propagation neural networks，BPNN）结合，进行冬小麦县级尺度产量预测。这些研究的结果表明，机器学习方法在一定程度上可以准确地预测作物产量。

收稿日期：2023-08-03 修订日期：2023-12-19
基金项目：井冈山农高区省级科技专项“揭榜挂帅”项目（20222-051244）；广东省基础与应用基础研究基金项目（2021A1515110442）
作者简介：裴杰，博士，助理教授，研究方向为农业遥感与粮食安全。Email：peij5@mail.sysu.edu.cn
※通信作者：方华军，博士，研究员，研究方向为农业生态与农业大数据。Email：fanghj@igsnrr.ac.cn

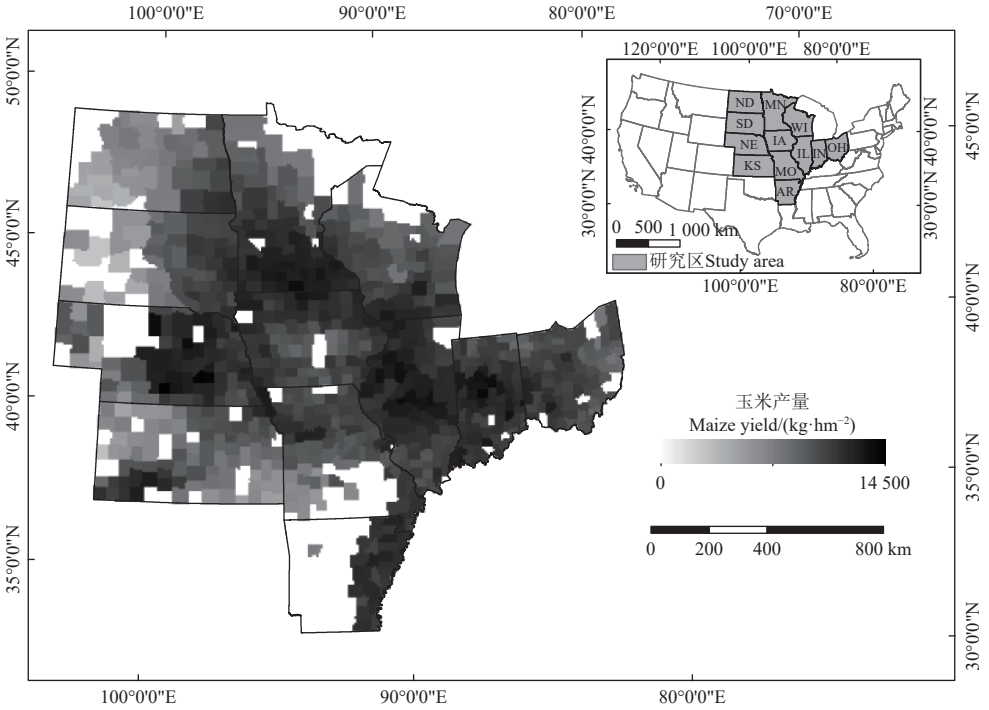
然而，作物产量预测具有空间属性，空间因素在预测中发挥着重要作用。以往研究大多建立于单一空间尺度上，很少考虑作物产量的空间变化。不考虑地理位置建模时，往往会在结果中引入不确定性[15-16]。大部分机器学习模型建立的均为全局模型，并未考虑数据的空间自相关性。而地理加权随机森林（geographically weighted random forest regression，GWRFR）在普通随机森林模型的基础上引入空间位置信息，并为每个位置的样本点建立一个局部模型，故在各种空间问题中表现优良，有着较强的稳定性[17-18]。此外，以往的大多数研究都是在全年或整个生长季的时间尺度上进行建模，对于产量的预测未能精细到玉米生长的各个物候阶段，无法给出玉米产量提前预测的最佳时间。

为此，本文预期实现以下两个主要目标：1）通过引入空间位置信息，构建地理加权随机森林模型，并与其他全局模型对比产量预测精度；2）在玉米的各个生长阶段中，确定最佳的产量预测时间窗口，从而实现产量早期预测。

1 材料与方法

1.1 研究区概况

本研究聚焦于美国中西部玉米带内的 12 个州，包括北达科他州（ND）、南达科他州（SD），堪萨斯州（KS）、内布拉斯加州（NE）、明尼苏达州（MN）、爱荷华州（IA）、威斯康星州（WI）、伊利诺伊州（IL）、印第安纳州（IN）、俄亥俄州（OH）、密苏里州（MO）和阿肯色州（AR）。美国玉米出口量约占全球的 15%，而玉米带上的玉米年产量占美国的 75% 以上，占全球的 36% 以上[5]。其中，2020 年美国中西部玉米带玉米产量分布如图 1 所示。



注：图中字母代表各州的简称。
Note: The letters in the figure represent abbreviations for each state.

图 1 2020 年美国中西部玉米带县级玉米产量分布
Fig.1 Distribution of county-level maize yield in the Corn Belt region of the midwestern United States in 2020

1.2 数据来源

1.2.1 产量及物候数据

本研究检索了研究区 2011—2020 年间的县级玉米产量记录，县级玉米产量数据来源于美国农业部（https://www.nass.usda.gov/），此外，该网站每年都会发布 1 期作物种植图层（cropland data layer，CDL）。为排除非玉米种植区域对本研究的影响，需要使用该图层提取玉米种植区域，并作为其他环境因子的掩膜。玉米物候数据来源于该网站的作物进展报告（crop progress report，CPR），该报告将玉米物候分为 7 个阶段，即：种植、出苗、抽丝、乳熟、腊熟、成熟和收获。该报告每周发布一期，分别统计了各种作物在当期处于各个物候阶段的比例。本研究将进展达到 20% 作为该物候阶段的开始时间，进展达到 80% 作为该物候阶段的结束时间。为了使各物候阶段时间连续，取相邻物候阶段的结束和开始时间的中间时刻作为最终两个物候阶段的分段点。对于整个生长季，其起始时间定义为玉米种植的开始时间，其结束时间定为收获期结束时间。最后得到美国玉米年度物候时间表（表 1）。研究中使用的行政区划数据来源于美国地质调查局（https://www.usgs.gov/），计算经纬度时仅计算各县玉米种植区域所在的中心位置，并用其生成每个样本点的空间权重。

表 1　2011—2020 年玉米物候期时间表

Table 1　Schedule of maize phenology period from 2011 to 2020

年份 Year	起止时间 Starting and ending time						
	种植 Planted	出苗 Emerged	抽丝 Silking	乳熟 Dough	腊熟 Dented	成熟 Mature	收获 Harvest
2011	05-01—05-18	05-18—06-22	06-22—07-31	07-31—08-17	08-17—09-07	09-07—10-02	10-02—10-30
2012	04-15—05-06	05-06—06-10	06-10—07-22	07-22—08-08	08-08—08-22	08-22—09-12	09-12—10-14
2013	05-05—05-22	05-22—06-26	06-26—08-04	08-04—08-28	08-28—09-15	09-15—10-09	10-09—11-10
2014	04-27—05-14	05-14—06-18	06-18—07-27	07-27—08-20	08-20—09-10	09-10—10-05	10-05—11-09
2015	04-26—05-13	05-13—06-21	06-21—07-26	07-26—08-19	08-19—09-06	09-06—09-30	09-30—11-01
2016	04-17—05-11	05-11—06-15	06-15—07-24	07-24—08-17	08-17—09-04	09-04—10-02	10-02—10-30
2017	04-23—05-14	05-14—06-21	06-21—07-30	07-30—08-16	08-16—09-10	09-10—10-08	10-08—11-12
2018	04-29—05-16	05-16—06-17	06-17—07-22	07-22—08-15	08-15—09-02	09-02—09-26	09-26—11-11
2019	05-05—05-29	05-29—06-30	06-30—08-04	08-04—08-25	08-25—09-18	09-18—10-16	10-16—11-24
2020	04-26—05-13	05-13—06-21	06-21—07-26	07-26—08-16	08-16—09-06	09-06—09-27	09-27—11-01

1.2.2　环境数据

本研究选取了气候、植被以及土壤 3 大类与玉米产量相关的环境数据，选取了其中 14 个环境因子作为模型的输入变量。环境因子的原始数据主要来源于 MODIS（moderate resolution imaging spectroradiometer）产品、PRISM（parameter-elevation regressions on independent slopes model）和 OpenLandMap 数据集，且均可通过谷歌地球引擎（google earth engine，GEE）获取，具体信息如表 2 所示。此外，由于研究的时间跨度较大，考虑到玉米种植技术随着时间而不断进步，故参考前人研究[14]将年份也作为自变量加入模型。

表 2　环境因子数据

Table 2　Environmental factors data

类型 Type	环境因子 Environmental factor	单位 Unit	TR/d	SR/m	数据来源 Data source
气候 Climate	平均降水量	mm	1	4 000	PRISM
	平均气温	℃	1	4 000	PRISM
	最高气温	℃	1	4 000	PRISM
	最低气温	℃	1	4 000	PRISM
	平均露点温度	℃	1	4 000	PRISM
	最小饱和水汽压差	hPa	1	4 000	PRISM
	最大饱和水汽压差	hPa	1	4 000	PRISM
	平均短波辐射	$W·m^{-2}$	1	1 000	Daymet
	蒸散发	$kg·m^{-2}$	8	500	MOD16A2
植被 Vegetation	归一化差异植被指数	-	16	1 000	MOD13A2
	增强型植被指数	-	16	1 000	MOD13A2
	总初级生产力	$kg·m^{-2}$	8	500	MOD17A2H
土壤 Soil	土壤有机碳	$g·kg^{-1}$	-	250	OpenLandMap
	土壤酸碱度 (pH)	-	-	250	

注：TR 和 SR 分别代表时间和空间分辨率。土壤有机碳以及土壤酸碱度均选择地下深度 30 cm。

Note: TR and SR represent temporal and spatial resolution, respectively. Soil organic carbon and soil pH were selected at a subsurface depth of 30 cm.

1.3　研究方法

1.3.1　数据预处理

首先根据统计的物候期时间表（表 1），对不同时间分辨率的原始环境数据在各个物候期内求取平均值，再对均值结果进行投影，统一所有数据的坐标系。随后使用玉米种植图层对各个环境因子进行掩膜和分区统计，得到各个县域每一年中不同生长阶段的均值结果。该流程均在 GEE 上完成。最后根据县域的联邦信息处理标准代码（federal information processing standard，FIPS）将统计的各县域玉米产量、空间位置以及环境因子关联起来，并对存在空值的数据条目进行删除。由于各个特征的量纲不同，本研究对原始数据进行了标准归一化处理，将数据转换成均值为 0，标准差为 1 的正态分布[19]，以此来提高模型的稳定性及精度[20]。最终在 7 个不同物候期以及整个生长季得到共 8 组数据，每组数据中都包含 7 995 个样本。

1.3.2　预测模型

本研究使用的产量预测模型包括多元线性回归（multiple linear regression，MLR）、支持向量回归（SVR）、梯度提升树（gradient boosting tree，GBT）、随机森林（RF）和地理加权随机森林（GWRFR）。5 种模型中，GWRFR 为局部模型，其他 4 种均为全局模型。其中 MLR 目的是建立多个自变量与一个因变量之间的线性关系，通过最小二乘法优化每个自变量对应的系数[21]。SVR 是一种监督学习模型，通过核方法处理非线性关系，将数据映射到高维空间以发现更复杂的模式[22]。GBT 是一种机器学习技术，是集成学习的一种形式，旨在通过结合多个弱学习器来创建一个强大的总体模型[23]。

RF 模型是一种基于构建大量决策树的集成学习算法[24]。其训练过程主要分为两个阶段：建树和投票。在建树阶段，RF 通过利用 bootstrap 重采样技术从原始训练样本集中随机抽取不同的数据子集，然后根据抽取的子集生成多个决策树并组成随机森林；而在投票阶段，随机森林通过对所有决策树的预测结果进行平均或多数投票来输出最终结果[25]。GWRFR 是一种基于 RF 开发的集成学习方法，用于改进非空间模型[18]。与传统的 RF 模型不同的是，GWRFR 不是建立一个全局模型，而是为每个样本点都构建一个局部模型，并根据空间位置来利用相邻样本点的观测信息，以此克服空间数据的“非平稳性”[26]，其原理与地理加权回归（geographically weighted regression，GWR）类似。为了描述 RF 与 GWRFR 这两种方法的区别，本研究采用回归方程的最简单形式：

$$Y_i = ax_i + e, i = 1:n \tag{1}$$

式中 Y_i 是第 i 次观测的因变量值，x_i 是第 i 个样本点，ax_i 是基于 RF 对 x_i 的非线性预测值，e 为模型误差项。但是该方程通常是使用所有数据构建的，并没有考虑样本的空间分布。但是在 GWRFR 中，本研究将式（1）扩展为

$$Y_i = a(u_i, v_i)x_i + e, i = 1:n \tag{2}$$

式中$a(u_i,v_i)x_i$是在位置i上通过空间权重对x_i校正的局部RF预测值，(u_i,v_i)是坐标。其意义是在每个样本点的位置上都建立一个子模型，并只考虑附近的观测结果[27]。其中，子模型的运行区域被称作邻域（也称作内核），邻域的半径被称为带宽[28]。邻域通常包含“自适应”和“固定”两种类型，前者是由n个最近的样本点定义，而后者由半径为带宽的圆定义[29-30]。

在本研究中，由于样本点的分布密度不均一，故使用自适应内核进行运算。此外，在使用GWRFR进行预测时可以设置全局模型与局部模型融合的比例，为了验证局部模型的精度是否优于全局模型，本研究将局部模型的比例设置为100%。

1.3.3 建模设计

本研究旨在探究GWRFR局部模型相较于其他传统全局模型是否能够显著提高产量预测精度。为此，本研究首先以玉米的整个生长季（即从种植至收获）作为时间窗口，对环境因子进行统计；并分别利用5种模型进行建模，对比GWRFR模型与其余4种全局模型在产量预测上的性能。这种对比分析将体现不同模型在处理相同数据集时的预测准确性，从而直观地评估GWRFR模型在作物产量预测领域的有效性。

为了进一步探究玉米产量预测与样本点空间位置之间的关系，本研究将在原有变量的基础上，在4种全局模型中加入经纬度坐标这一变量，重新进行产量建模。随后，将加入经纬度变量的全局模型建模结果与GWRFR模型的精度进行对比，选取精度最高的模型作为最优模型。

最后，在得到最优模型后，为了实现玉米产量早期预测，并获取最佳预测时间，本研究以不同物候期的起始和结束时间作为时间窗口，提取了不同物候期中的环境因子。针对不同物候期建立模型时，本研究采用了两种不同的策略：第一种策略是物候期累积变量建模，即将当前物候阶段与前序物候期的环境变量累积一起输入模型，例如在建立抽丝期的模型时，会将种植期、出苗期和抽丝期这3个物候期的所有变量共同输入模型进行产量建模。该策略可以捕捉整个生长周期内各因素对产量的综合影响，但由于早期的物候期数据同样会影响到后期的预测结果，故也可能会掩盖某些特定物候期的特定影响；第二种策略是单物候期变量建模，即在建立每个物候期的模型时，仅使用此物候期的变量进行建模。此方法专注于单独一个物候期内的变量，忽略其他物候期的数据。这种策略能更精确地分析特定物候期对产量的影响，但可能会忽略不同物候期间的交互作用和累积效应。两种方式各有侧重，最终都通过不同物候期模型的精度变化曲线获取最佳预测时间。

在进行模型之间的比较时，本研究将10 a内的所有数据随机化处理，并按照7：3的比例划分为训练集和测试集。使用训练集构建模型，并采用网格搜索算法结合十折交叉验证的方法寻找模型的最优超参数[31]，最后将最优的模型应用于测试集，进行精度验证。此外，本研究还采用逐年精度验证的方法进一步验证模型的可靠性，即从10 a数据中选取9 a的数据作为训练集建立模型，剩下的1 a数据作为测试集验证模型精度，如此重复10次直到得到每一年的验证精度。本研究采用均方根误差（root mean squared error，RMSE）以及决定系数（coefficient of determination，R^2）等指标来评估模型的精度以及泛化能力[32]。

2 结果与分析

2.1 全局与局部模型精度对比

以整个生长季作为时间窗口，将不包含经纬度的所有变量输入模型，以对比全局模型与局部模型在输入变量相同时的精度（表3）。结果表明，地理加权随机森林（GWRFR）的精度高于其他4种全局模型。其中，多元线性回归（MLR）模型的精度相较于其他模型明显偏低。这是由于该模型在处理仅涉及线性关系的场景中较有效，但其难以捕捉作物产量与环境因子之间的非线性关系。支持向量回归SVR在处理小至中等规模数据时表现良好，但在处理非常大规模的数据时一般不如基于树的方法，所以其模型精度要略低于梯度提升树GBT、随机森林RF和GWRFR模型。相较于GBT，RF能更好地处理大量特征，且对于特征选择的敏感性低，这使其在处理包含多个相关变量的复杂数据集时效果更佳。GWRFR在RF的基础上引入了地理信息，提供了额外的、对产量预测至关重要的空间上下文信息。通过为不同地理位置创建独立的模型，GWRFR揭示了隐藏的空间模式和趋势，可以适应各地区的独特特性和条件。这种空间维度信息的加入是GWRFR在性能上超越其他全局模型的关键。

表3 不同模型建模精度

Table 3 Accuracy of different models

模型 Model	加入经纬度 Add coordinate	R^2	RMSE/(kg·hm^{-2})
多元线性回归 Multiple linear regression (MLR)	否	0.62	1 510.63
	是	0.64	1 435.20
支持向量回归 Support vector regression (SVM)	否	0.81	1 053.11
	是	0.84	930.67
梯度提升树 Gradient boosting tree (GBT)	否	0.82	1 039.91
	是	0.85	911.59
随机森林 Random forest (RF)	否	0.83	994.75
	是	0.85	890.88
地理加权随机森林 Geographically weighted RF regression (GWRFR)	-	0.87	864.21

注：R^2为决定系数；RMSE为均方根误差，下同。

Note: R^2 is coefficient of determination, and RMSE is root mean squared error, same below.

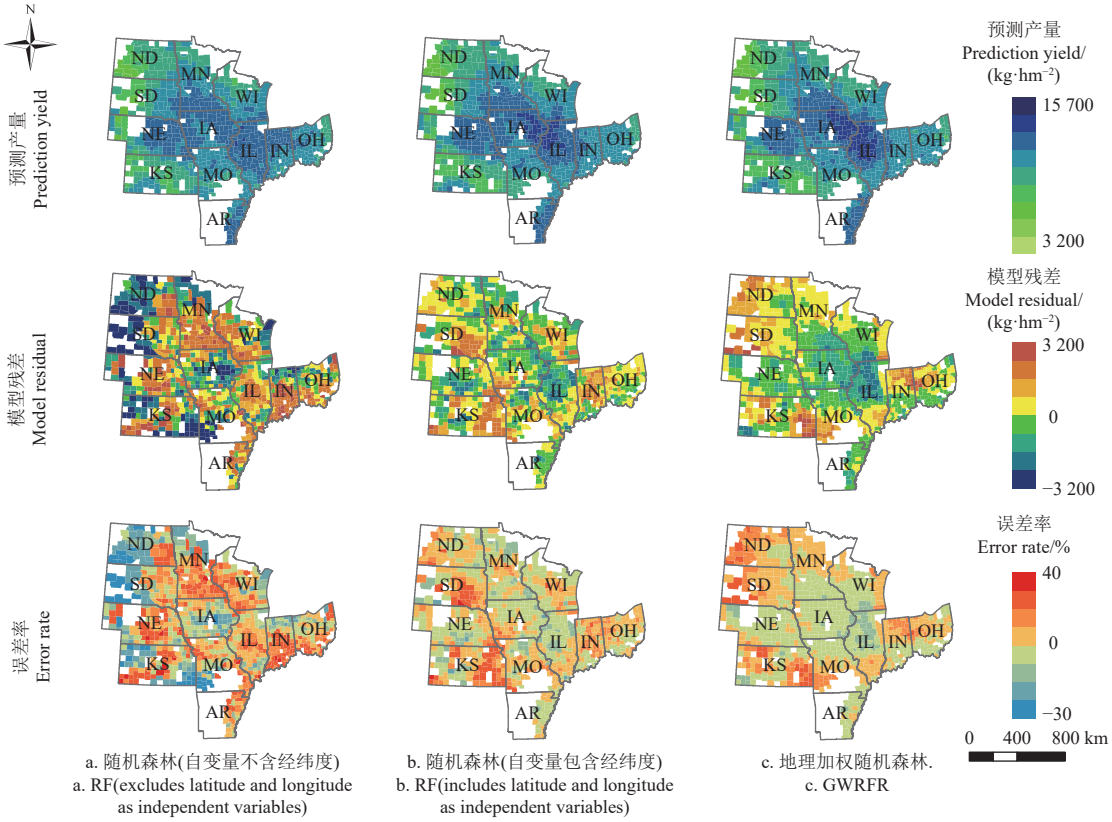
为了进一步比较局部模型GWRFR与全局模型在产量预测中的可靠性，本研究使用GWRFR与最优全局模型RF进行逐年的精度验证，结果如表4所示。在10年中，除2012年与2014年外，GWRFR模型的精度均高于RF模型。逐年精度验证进一步增大了RF模型与GWRFR之间的精度差距，表明GWRFR模型具有更强的泛化能力。

表 4　随机森林与地理加权随机森林逐年验证精度

Table 4　Annual validation accuracy of RF and GWRFR

年份 Year	模型 Model	R^2	RMSE/(kg·hm^{-2})	年份 Year	模型 Model	R^2	RMSE/(kg·hm^{-2})
2011	RF	0.57	1 404.57	2016	RF	0.56	1 247.04
	GWRFR	0.61	1 327.37		GWRFR	0.61	1 173.61
2012	RF	0.18	2 424.42	2017	RF	0.51	1 482.39
	GWRFR	0.08	2 559.98		GWRFR	0.60	1 336.16
2013	RF	0.57	1 299.13	2018	RF	0.58	1 445.99
	GWRFR	0.69	1 115.87		GWRFR	0.63	1 356.87
2014	RF	0.58	1 224.45	2019	RF	0.25	1 540.13
	GWRFR	0.53	1 285.95		GWRFR	0.55	1 197.46
2015	RF	0.40	1 435.95	2020	RF	0.62	1 216.92
	GWRFR	0.44	1 396.41		GWRFR	0.66	1 151.02

为了深入研究玉米产量预测与样本点空间位置之间的关系，本研究将经纬度坐标作为变量加入全局模型，以对比包含地理位置信息的全局模型与局部模型的精度（表 3）。结果表明，全局模型在加入经纬度坐标这一变量后，模型精度获得了提升，其中 RF 的精度仍是全局模型中最高的（R^2=0.85，RMSE=890.88 kg/hm^2），但仍低于 GWRFR 的精度（R^2=0.87，RMSE=864.21 kg/hm^2）。因此，对于作物产量预测来说，加入空间位置信息能有效提升产量预测精度。并且，局部模型由于可以考虑不同地理区域的空间异质性，因此能更有效地捕捉作物产量随环境因子的空间变化模式，从而更准确地预测产量。图 2 以 2020 年为例，展示了 3 种方案产量预测结果以及误差的空间分布。相较于前两种全局模型，GWRFR 预测的产量空间分布更加集中，一致性更强，这是由于该方法能够更好地捕捉数据的空间信息。分析预测残差分布图发现，3 种方案在研究区西北部、南部均出现了较大的预测误差。研究区西北部和南部地区是低产量区域，此区域主要呈现产量高估现象，这是机器学习算法中常见的问题[33-35]。造成该问题的原因是极端低值数据样本量通常较少，这使得机器学习模型对这些极端值的有效预测变得更加困难。尽管 GWRFR 模型并不能完全突破机器学习框架的限制，但相较于传统的随机森林模型，经过改良后的 GWRFR 显然对低值高估问题有一定的改善。



注：模型残差等于预测值减真实值，误差率等于模型残差除以真实产量。

Note: The model residual is equal to the predicted value minus the true value. The error rate is equal to the model residual divided by the real yield.

图 2　2020 年产量预测以及误差的空间分布

Fig.2　Spatial distribution of yield prediction and error in 2020

2.2　玉米产量最佳预测时间分析

在确定了局部模型 GWRFR 优于其他全局模型后，本研究使用此模型，分别利用物候期累积变量建模和单物候期建模两种不同的建模策略获取玉米产量的最佳预测时间（表 5）。在物候期累积建模中，当自变量随物候期逐渐累积，其预测精度在玉米生长前期逐步提升，在第 4 个物候期（乳熟期）精度达到最高（R^2=0.90，RMSE=748.39 kg/hm^2）。但再增加物候期，精度将趋于稳定并略有下降。这一现象是因为在作物生长初期，输入变量的增加丰富了作物生长信息，提高了模型的精度；但在乳熟期之后，新增加的变量对产量预测的贡献却相对有限，这可能导致模型在学习早期阶段关键特征时受到干扰，从而使模型预测精度呈现稳定或轻微下降的趋势。而对于单物候期建模，其预测精度在第 1～3 个物候

期（种植—抽丝）不断提升，在抽丝期达到最高（R^2=0.88，RMSE=827.85 kg/hm^2），之后不断下降。这是由于在抽丝期后，玉米逐渐成熟，对环境变化的敏感程度不如前 3 个物候期强烈，导致环境因子对产量预测的效果逐渐下降[36-37]。综上所述，在本研究中，玉米产量最早可以在乳熟期进行预测，最佳预测时间可以提前至收获前 2～3 个月。此外，在玉米的整个生长季内，抽丝期对玉米产量预测的影响最大。

对比累积变量建模的最佳结果（种植—乳熟期）、单物候期变量建模的最佳结果（抽丝期）与全生长季建模得到的结果（种植—收获）的空间分布，如图 3 所示。3 个模型所得玉米产量预测结果的空间分布都比较集中。同样地，在研究区的西部外围以及中部地区出现了较大的预测残差。3 个模型的预测 R^2 分别为 0.90、0.88、0.87，RMSE 分别为 748.39、827.85、864.21 kg/hm^2。3 种方案的决定系数 R^2 相差约 0.01，甚至早期预测的精度优于以整个生长季作为时间窗口的预测结果，这表明在抽丝期或乳熟期进行玉米产量的早期预测是准确可靠的。

表 5 建模精度随作物物候的变化

Table 5 Change of modeling accuracy with crop phenology

建模策略 Modeling strategy	物候期 Phenological period	R^2	RMSE/(kg·hm^{-2})
物候期累积建模 Phenological cumulative modeling	种植	0.84	940.89
	出苗	0.86	872.00
	抽丝	0.88	802.60
	乳熟	0.90	748.39
	腊熟	0.89	762.86
	成熟	0.90	756.78
	收获	0.89	777.74
单物候期建模 Single phenological stage modeling	种植	0.85	915.17
	出苗	0.85	929.79
	抽丝	0.88	827.85
	乳熟	0.87	852.18
	腊熟	0.87	848.06
	成熟	0.85	924.69
	收获	0.85	926.06

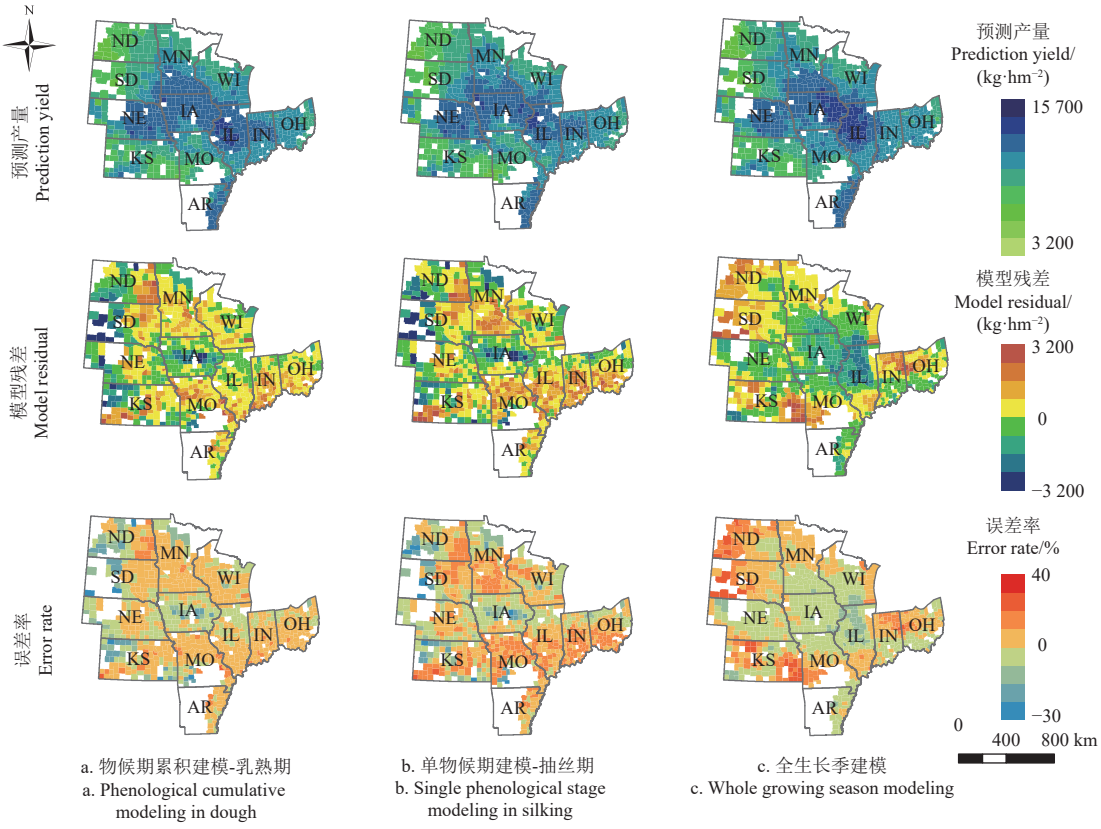


图 3 2020 年最佳物候期与生长季产量预测及误差空间分布

Fig.3 Spatial distribution of yield prediction and error for optimal phenological period and growing season in 2020

3 讨 论

本文分析结果表明，与其他常见模型相比，GWRFR 模型在产量预测中表现更好。这是由于空间位置对于玉米产量预测有着重要的作用，相比于传统的全局模型，考虑空间异质性的 GWRFR 局部模型的预测精度以及泛用性更高。KHAN 等[38] 使用了 5 组不同的特征集合并结合 GWRFR 等 6 种流行的机器学习算法对美国玉米带地区玉米产量进行预测，结果发现在 5 个特征集合中，GWRFR 模型精度均高于其他模型。这与本研究的结果一致。对于地理数据而言，利用空间加权的方法考虑空间异质性比简单地加入经纬度坐标作为变量有着更好的表现，在作物产量预测方面精度也更高。此外，在玉米产量预测的空间分布上，GWRFR 模型的预测结果分布更加集中，误差高值区域更小，表明 GWRFR 比传统的非空间模型能够更好地克服地理数据的非平稳性。尽管在本研究中，GWRFR 在美国县级玉米产量预测方面的性能要优于全局模型，但未来依然需要更多的研究来评

估其在不同空间尺度上预测其他地区、其他作物产量的有效性。

物候期累积建模的结果表明，玉米产量预测的最佳物候期为乳熟期，这是因为随着作物的生长，环境对作物影响的累积效应在乳熟期达到峰值，此后，环境对产量的影响相对前期而言逐渐减弱。而在单物候期建模中，研究发现抽丝期是对玉米产量影响最大的一个阶段，这说明前者的结论是可靠的；其次，对于物候期累积建模来说，乳熟期之后的建模精度保持较小的波动，也从侧面印证了乳熟期是玉米产量最佳提前预测时间这一结论的准确性。另外，此前有部分学者也曾开展过相似的研究，例如：MENG 等[39] 基于遥感光谱指数和线性回归方法预测了中国东北地区的玉米产量，研究结果表明，最佳提前预测时间为播种后 55～60d，即在抽丝期快结束时。LI 等[11] 基于多源遥感数据和随机森林方法预测了中国 3 种主要作物的产量，其中玉米的最佳提前预测时间约为收获前的 1～2 个月。其研究结论与本研究结果有细微的差异，这可能是由于研究区地理位置、玉米品种以及建模方法不同所导致的。

4　结　论

本研究使用了多源遥感数据并结合多种传统机器学习模型以及地理加权随机森林（GWRFR）模型对美国县级玉米产量进行建模，探讨全局与局部模型在玉米产量预测方面的性能差异，进而将 GWRFR 模型应用于玉米的各个物候期，获取了最佳的玉米产量提前预测时间。主要结论如下：

1）相较于传统的全局模型，结合空间因素构建的局部模型不仅在预测精度上有一定的提升，而且其产量预测结果的空间分布也更加集中，能够较好地克服空间数据的非平稳性。

2）美国玉米产量最早可以在乳熟期进行准确预测，最佳预测时间可以提前至收获前 2～3 个月。而在玉米整个生长阶段中，抽丝期对玉米产量预测的影响最大。

[参　考　文　献]

[1] 江昊. 数据驱动的玉米估产方法和时空特性分析研究 [D]. 杭州：浙江大学，2022.
JIANG Hao. Data Driven Corn Yield Estimation Methods and Spatiotemporal Characteristic Analysis [D]. Hangzhou: Zhejiang University, 2022. (in Chinese with English abstract)

[2] KHAKI S, WANG L. Crop yield prediction using deep neural networks[J]. Frontiers in Plant Science, 2019, 10: 621.

[3] PANTAZI X E, MOSHOU D, ALEXANDRIDIS T, et al. Wheat yield prediction using machine learning and advanced sensing techniques[J]. Computers and Electronics in Agriculture, 2016, 121: 57-65.

[4] 王汇涵，张泽，康孝岩，等. 基于 Sentinel-2A 的棉花种植面积提取及产量预测[J]. 农业工程学报，2022，38(9)：205-214.
WANG Huihan, ZHANG Ze, KANG Xiaoyan, et al. Cotton planting area extraction and yield prediction based on Sentinel-2A[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(9): 205-214. (in Chinese with English abstract)

[5] RANUM P, PENA-ROSAS J P, GARCIA-CASAL M N. Global maize production, utilization, and consumption[J]. Annals of the New York Academy of Sciences, 2014, 1312: 105-112.

[6] 彭丽. 基于 MODIS 和气象数据的陕西省小麦与玉米产量估算模型研究 [D]. 杭州：浙江大学，2014.
PENG Li. Wheat and Maize Yield Model Estimation Based on MODIS and Meteorological Data in Shaanxi Province [D]. Hangzhou: Zhejiang University, 2014. (in Chinese with English abstract)

[7] FENG P, WANG B, LIU D, et al. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique[J]. Agricultural and Forest Meteorology, 2020, 285: 107922.

[8] ARCHONTOULIS S V, CASTELLANO M J, LICHT M A, et al. Predicting crop yields and soil‐plant nitrogen dynamics in the US Corn Belt[J]. Crop Science, 2020, 60(2): 721-738.

[9] 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 Tengcong, 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 (Transactions of the CSAE), 2017, 33(19): 147-155. (in Chinese with English abstract)

[10] LEROUX L, CASTETS M, BARON C, et al. Maize yield estimation in West Africa from crop process-induced combinations of multi-domain remote sensing indices[J]. European Journal of Agronomy, 2019, 108: 11-26.

[11] LI L, WANG B, FENG P, et al. Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China[J]. Agricultural and Forest Meteorology, 2021, 308: 108558.

[12] 严海军，卓越，李茂娜，等. 基于机器学习和无人机多光谱遥感的苜蓿产量预测[J]. 农业工程学报，2022，38(11)：64-71.
YAN Haijun, ZHUO Yue, LI Maona, et al. Alfalfa yield prediction using machine learning and UAV multispectral remote sensing[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(11): 64-71. (in Chinese with English abstract)

[13] 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.
LIU Junming, HE Xiaotong, WANG Pengxin, et al. Early prediction of winter wheat yield with long time series meteorological data and random forest method[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2019, 35(6): 158-166. (in Chinese with English abstract)

[14] 孙少杰，吴门新，庄立伟，等. 基于 CNN 卷积神经网络和

BP 神经网络的冬小麦县级产量预测[J]. 农业工程学报，2022，38(11)：151-160.
SUN Shaojie, WU Menxin, ZHUANG Liwei, et al. Forecasting winter wheat yield at county level using CNN and BP neural networks[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(11): 151-160. (in Chinese with English abstract)

[15] OSHAN T M, SMITH J P, FOTHERINGHAM A S. Targeting the spatial context of obesity determinants via multiscale geographically weighted regression[J]. International Journal of Health Geographics, 2020, 19(1): 1-17.

[16] 覃文忠. 地理加权回归基本理论与应用研究 [D]. 上海：同济大学，2007.
QIN Wenzhong. The Basic Theoretics and Application Research on Geographical Weighted Regression [D]. Shanghai: Tongji University, 2007. (in Chinese with English abstract)

[17] LUO Y, YAN J, MCCLURE S C, et al. Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model[J]. Environmental Science and Pollution Research, 2022, 29(22): 33205-33217.

[18] SANTOS F, GRAW V, BONILLA S. A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon[J]. PLoS One, 2019, 14(12): e0226224.

[19] PANDEY A, JAIN A. Comparative analysis of KNN algorithm using various normalization techniques[J]. International Journal of Computer Network and Information Security, 2017, 9(11): 36-42.

[20] WAN X. Influence of feature scaling on convergence of gradient iterative algorithm[C]. Bristol, US: IOP Publishing, 2019.

[21] 冷建飞，高旭，朱嘉平. 多元线性回归统计预测模型的应用[J]. 统计与决策，2016(7)：82-85.

[22] DEWI C, CHEN R C. Random forest and support vector machine on features selection for regression analysis[J]. International Journal of Innovative Computing, Information and Control, 2019, 15(6): 2027-2037.

[23] 张宏鸣，刘雯，韩文霆，等. 基于梯度提升树算法的夏玉米叶面积指数反演[J]. 农业机械学报，2019，50(5)：251-259.
ZHANG Hongming, LIU Wen, HAN Wenting, et al. Inversion of summer maize leaf area index based on gradient boosting decision tree algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(5): 251-259. (in Chinese with English abstract)

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

[25] HASTIE T, TIBSHIRANI R, FRIEDMAN J, et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction [M]. New York, USA: Springer, 2009.

[26] FOTHERINGHAM A S, CHARLTON M, BRUNSDON C. The geography of parameter space: An investigation of spatial non-stationarity[J]. International Journal of Geographical Information Systems, 1996, 10(5): 605-627.

[27] GEORGANOS S, GRIPPA T, GADIAGA A N, et al. Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling[J]. Geocarto International, 2021, 36(2): 121-136.

[28] BRUNSDON C, FOTHERINGHAM S, CHARLTON M. Geographically weighted regression[J]. Journal of the Royal Statistical Society:Series D (The Statistician), 1998, 47(3): 431-443.

[29] FOTHERINGHAM A S, BRUNSDON C, CHARLTON M. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships [M]. New York, USA: John Wiley & Sons, 2003.

[30] KALOGIROU S. Destination choice of athenians: An application of geographically weighted versions of standard and zero inflated poisson spatial interaction models[J]. Geographical Analysis, 2016, 48(2): 191-230.

[31] SHAHHOSSEINI M, HU G, HUBER I, et al. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt[J]. Scientific Reports, 2021, 11(1): 1606.

[32] JEONG J H, RESOP J P, MUELLER N D, et al. Random forests for global and regional crop yield predictions[J]. PLoS One, 2016, 11(6): e0156571.

[33] LENG G, HALL J W. Predicting spatial and temporal variability in crop yields: An inter-comparison of machine learning, regression and process-based models[J]. Environmental Research Letters, 2020, 15(4): 044027.

[34] YIN X, LENG G, YU L. Disentangling the separate and confounding effects of temperature and precipitation on global maize yield using machine learning, statistical and process crop models[J]. Environmental Research Letters, 2022, 17(4): 044036.

[35] 汪静平，吴小丹，马杜娟，等. 基于机器学习的遥感反演：不确定性因素分析[J]. 遥感学报，2023，27(3)：790-801
WANG Jingping, WU Xiaodan, MA Dujuan, et al. Remote sensing retrieval based on machine learning algorithm: Uncertainty analysis[J]. National Remote Sensing Bulletin, 2023, 27(3): 790-801. (in Chinese with English abstract)

[36] LIU W, TOLLENAAR M, STEWART G, et al. Response of corn grain yield to spatial and temporal variability in emergence[J]. Crop Science, 2004, 44(3): 847-854.

[37] 王亚许，孙洪泉，吕娟，等. 基于 APSIM 模型的春玉米生育期旱灾损失敏感性定量分析[J]. 灾害学，2021，36(2)：30-36.
WANG Yaxu, SUN Hongquan, LV Juan, et al. Quantitative analysis of the sensitivity of spring maize to drought in the growth period based on APSIM model[J]. Journal of Catastrophology, 2021, 36(2): 30-36. (in Chinese with English abstract)

[38] KHAN S N, LI D, MAIMAITIJIANG M. A geographically weighted random forest approach to predict corn yield in the US Corn Belt[J]. Remote Sensing, 2022, 14(12): 2843.

[39] MENG W, TAO F L, SHI W J. Corn yield forecasting in northeast China using remotely sensed spectral indices and crop phenology metrics[J]. Journal of Integrative Agriculture, 2014, 13(7): 1538-1545.

Early prediction of maize yield by integrating GWRFR and crop phenological information

PEI Jie[1,2], TAN Shaofeng[1], GUO Han[1], LIU Yibo[1], FANG Huajun[3,4※]

(1. *School of Geospatial Engineering and Science*, *Sun Yat-sen University*, *Zhuhai* 519082, *China*;　2. *Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical Area of South China*, *Ministry of Natural Resources*, *Zhuhai* 519082, *China*;　3. *Key Laboratory of Ecosystem Network Observation and Modeling*, *Institute of Geographical Sciences and Natural Resources Research*, *Chinese Academy of Sciences*, *Beijing* 100101, *China*;　4. *The Zhongke-Ji'an Institute for Eco-Environmental Sciences*, *Ji'an* 343000, *China*)

Abstract: The precise estimation of crop yields is essential for global food security, particularly in the face of challenges like climate change, population growth, and food distribution inequalities. Despite the widespread use of machine learning techniques combined with remote sensing data for large-scale yield prediction, the integration of crop spatial position information and local models remains underexplored. This is particularly significant given the spatial nature of crop yield prediction, where spatial factors are highly influential. Previous studies, predominantly conducted on an annual or full-growth season basis, have not provided precise predictions for each phenological stage of maize growth. Consequently, these studies fall short in pinpointing the most effective prediction time for maize yield and understanding the impact of environmental factors at each stage. This research delves into two key questions: 1) Does the inclusion of spatial location information in the geographic weighted random forest (GWRFR) model improve yield prediction accuracy over the traditional random forest model? 2) Among different phenological stages of maize, which stage provides the optimal window for yield prediction? To address these issues, this study employed multi-source remote sensing data in conjunction with machine learning algorithms, and predicted maize yield at the county level in the United States. This study investigated the relationship between yield prediction and the spatial location of sample points, assessing the relevance of including latitude and longitude as independent variables. Further, the study introduced the local GWRFR model for maize yield prediction and compared its modeling performance with the global random forest (RF) model. In addition, the study examined two methodological approaches for determining the best prediction time. The first approach, referred to as the accumulated environmental variables (AEV) approach, integrated data from various phenological periods. The second approach, known as the current stage variables (CSV) approach, used data exclusively from the specific growth stage under analysis. The seven key growth stages of maize included planted, emerged, silking, dough, dent, mature and harvest, providing a comprehensive view of the crop's lifecycle. Through a comprehensive evaluation of the results from both schemes, this study identified the optimal prediction time for maize yield. The findings indicate that incorporating latitude and longitude into the model enhanced yield prediction accuracy. Without these spatial factors, the RF model achieved an coefficient of determination (R^2) of 0.83 and root mean squared error (RMSE) of 994.75 kg/hm^2, while including them improved these metrics to an R^2 of 0.85 and RMSE of 890.88 kg/hm^2. This provides preliminary evidence that including spatial factors can enhance maize yield prediction accuracy. Moreover, the local GWRFR model further improved prediction accuracy (R^2=0.87, RMSE=864.21 kg/hm^2), outperforming the traditional RF model and effectively addressing the non-stationarity of spatial data. In terms of optimal prediction time, the scheme where the environmental variables accumulate over phenological stages showed increasing accuracy from the first stage (planted) up to the fourth stage (dough), peaking at R^2=0.90 and RMSE of 748.39 kg/hm^2, and then stabilized. In contrast, the scheme utilizing only current stage variables improved accuracy from the first stage up to the third stage (silking), reaching its peak (R^2=0.88, RMSE=827.85 kg/hm^2) before decreasing. This suggests the best prediction time was around dough stage, approximately 2-3 months before harvest. Additionally, the strong correlation observed between early prediction results and those covering the entire growth season underscores the reliability of maize yield predictions made during the dough stages. In conclusion, this study introduces a novel method for large-scale crop yield prediction, integrating spatial data and phenological stages with advanced modeling techniques. The findings significantly contribute to enhancing food security and stabilizing the global food supply chain. This research not only provides critical insights for agricultural practices but also sets a foundation for future studies in crop yield prediction, potentially extending to other crops and regions, and incorporating a broader range of environmental factors.

Keywords: yield; prediction; remote sensing; machine learning; crop phenology