doi:10.6041/j.issn.1000-1298.2019.07.026

基于随机森林回归的玉米单产估测

王鹏新[1,2]　齐　璇[1,2]　李　俐[2,3]　王　蕾[1,2]　许连香[1,2]

(1. 中国农业大学信息与电气工程学院, 北京 100083; 2. 农业农村部农业灾害遥感重点实验室, 北京 100083;
3. 中国农业大学土地科学与技术学院, 北京 100083)

摘要: 为了提高玉米单产估测精度,以河北省中部平原为研究区域,以条件植被温度指数(VTCI)和上包络线 S-G 滤波的叶面积指数(LAI)为特征变量,通过随机森林回归确定玉米主要生育时期 VTCI 和 LAI 的权重,构建加权 VTCI 和 LAI 与玉米单产的单变量和双变量估产模型。结果表明,基于随机森林回归的双变量估产模型精度最高($R^2=0.303$),达极显著水平($P<0.001$)。将随机森林回归双变量估产模型用于研究区域 2012 年各县(区)玉米单产估测,结果表明,53 个县(区)玉米估测单产与实际单产的平均相对误差为 9.85%,均方根误差为 824.77 kg/hm^2,模型精度较高。基于随机森林回归双变量估产模型逐像素估测研究区域 2010—2018 年玉米单产,结果表明,玉米单产在空间上的分布特征为西部地区最高、北部和南部次之、东部地区最低,年际间的分布特征为在波动中呈先减少后增加的趋势。

关键词: 玉米; 估产; 条件植被温度指数; 叶面积指数; 随机森林回归

中图分类号: TP79　　**文献标识码**: A　　**文章编号**: 1000-1298(2019)07-0237-09

Estimation of Maize Yield Based on Random Forest Regression

WANG Pengxin[1,2]　QI Xuan[1,2]　LI Li[2,3]　WANG Lei[1,2]　XU Lianxiang[1,2]

(1. *College of Information and Electrical Engineering, China Agricultural University, Beijing* 100083, *China*
2. *Key Laboratory of Remote Sensing for Agri-Hazards, Ministry of Agriculture and Rural Affairs, Beijing* 100083, *China*
3. *College of Land Science and Technology, China Agricultural University, Beijing* 100083, *China*)

Abstract: Dynamic monitoring of crop growth and accurate estimation of crop yield can provide effective support for agricultural operators' field management and national food policy formulation. In order to improve the estimation accuracy of maize yield, a study was carried out in central plain of Hebei Province, including Baoding City, Shijiazhuang City, Cangzhou City, Hengshui City and Langfang City, from 2010 to 2018. The experiment was characterized by remotely sensed vegetation temperature condition index (VTCI) and Savitzky-Golay filtered leaf area index (LAI), which were closely related to maize growth and yield. Because the effects of water stress on maize yield at different growth stages were different, the weights of VTCI and LAI in the main growth stages (seedling-jointing, jointing-booting, booting-milking, milking-mature) of maize were determined by using the random forest regression method. The results showed that the weights based on the random forest regression were consistent with the actual growth of maize. Based on the determined weights, the weighted VTCI and LAI at the main growth stages of maize in each county (district) were calculated, and the univariate and bivariate estimation models of weighted VTCI and LAI with maize yield in 2010—2016 (except 2012) were constructed. The results showed that the accuracy of the bivariate estimation model ($R^2=0.303$) was higher than that of the univariate estimation models, and the bivariate model reached a very significant level ($P<0.001$), indicating that maize yield was related to VTCI and LAI. In summary, the bivariate estimation model based on the random forest regression had the highest accuracy. The bivariate estimation model based on the random forest regression was used to estimate the maize yield in each county (district) of the study area in 2012. The results showed that the average relative error between estimated yield and actual yield of 53 counties (districts) was 9.85%, and that of 31 counties

收稿日期: 2019-01-08　修回日期: 2019-01-30

基金项目: 国家重点研发计划项目(2016YFD0300603-3)

作者简介: 王鹏新(1965—),男,教授,博士,主要从事定量遥感及其在农业中的应用研究,E-mail: wangpx@cau.edu.cn

(districts) were below 10%, 7 counties (districts) were between 10% and 15%, 15 counties (districts) were more than 15% and the root mean square error was 824.77 kg/hm^2. In order to further verify the accuracy of the bivariate estimation model, a linear regression analysis model between actual yield and estimated yield of maize in 2012 was established. It could be seen that there was a significant positive correlation between estimated yield and actual yield ($P<0.001$) and R^2 reached 0.540, further indicating that the accuracy of the bivariate estimation model based on random forest regression was high. The bivariate estimation model based on the random forest regression was used to estimate the yield of maize in the region from 2010 to 2018. The results showed that the spatial distribution of maize yield was the highest in the western region of the plain, the next was in the north and south regions, and the lowest was in the eastern region. The distribution in time was characterized by a tendency to decrease first in the fluctuations and then increase. This was consistent with the actual spatial and temporal distribution characteristics of maize yield. The research result can provide reference for maize growth monitoring and yield estimation.

Key words: maize; estimation of yield; vegetation temperature condition index; leaf area index; random forest regression

0 引言

作物长势的动态监测及产量的准确估测,能够为农业经营者的田间管理和国家粮食政策的制定提供有效支撑[1-2]。近年来,随着遥感技术的迅速发展,大范围、多维空间的作物长势监测和产量估测成为可能。目前,经验回归模型是作物产量估测的常用方法之一[3]。

经验回归模型通常选取与作物产量密切相关的特征参数进行估产。在此类研究中,植被指数(Vegetation index, VI)应用广泛[4]。任建强等[5]以美国玉米为研究对象,以各州为估产区,通过筛选的归一化植被指数(Normalized difference vegetation index, NDVI)与玉米单产间的最佳估产模型对2011年各州玉米单产进行了估算,并推算全国玉米单产,结果表明,全国玉米单产的相对误差仅为2.12%。王恺宁等[6]选取Landsat 8 OLI卫星遥感数据,计算冬小麦灌浆期归一化植被指数、比值植被指数(Ratio vegetation index, RVI)、绿度植被指数(Greenness vegetation index, GVI)和增强植被指数(Enhanced vegetation index, EVI)4种植被指数,并与冬小麦单产建立单植被指数和多植被指数的神经网络和SVM模型,结果表明,多植被指数SVM模型的估产精度高于神经网络模型。LIAQAT等[7]以巴基斯坦整个印度河流域为研究区域,通过多种植被指数,如土壤调整植被指数(Soil adjusted vegetation index, SAVI)和改良土壤调整植被指数(Modified soil adjusted vegetation index, MSAVI)等,与小麦单产建立逐步回归模型,结果表明SAVI与小麦单产的决定系数R^2和Pearson相关系数分别为0.74和0.88。然而,作物单产除与植被指数相关外,还与土壤含水率和生长状态密切相关[8]。因此,可通过综合作物生长过程中的水分胁迫指标和生长状态指标提高作物单产估测精度。其中,条件植被温度指数(Vegetation temperature condition index, VTCI)是基于归一化植被指数和地表温度(Land surface temperature, LST)的散点图呈三角形的基础上提出的[9],可用于定量化地表征作物水分胁迫信息,并已成功应用于陕西省关中地区干旱监测及冬小麦单产估测预测等[10-12]。叶面积指数(Leaf area index, LAI)可表征植物的生长状态和光合作用能力,是作物长势监测及单产估测的重要指标[13]。此外,不同生育时期发生水分胁迫对作物单产的影响程度不同[14],可通过赋予不同生育时期特征变量不同的权重,构建综合特征参数进行作物单产估测以提高估测精度。王鹏新等[15]利用重采样粒子滤波算法同化VTCI和LAI,并基于组合熵的方法构建加权VTCI和LAI与冬小麦单产的线性回归模型,结果表明不同管理模式下影响冬小麦单产的主要因子不同。

随机森林(Random forest, RF)回归模型是一种流行的机器学习模型,具有抗过拟合和预测精度高的特点[16-17]。应用随机森林回归估测作物单产(尤其是通过综合指数估测单产)的研究相对较少。因此,本文以河北省中部平原地区为研究区域,选取条件植被温度指数和叶面积指数为特征变量,通过随机森林回归算法获取玉米主要生育时期各个特征变量的权重,进而构建加权特征变量与玉米单产间的回归模型,以期为作物长势监测及单产估测提供新思路。

1 材料与方法

1.1 研究区概况

河北省中部平原处于东经114°32′~117°36′,

北纬 36°57′～39°50′之间(图 1),包括石家庄市、保定市、廊坊市、衡水市和沧州市的部分或全部地区,包含 53 个县(区)。该区域属暖温带大陆性季风气候,四季分明,降水集中,是华北平原重要的农业生产区之一。该地区年降水量在 350～700 mm 之间,且时空分布不均,降水主要集中在夏季,占全年的 65%～70%,降水量由南向北逐渐减少。冬小麦-夏玉米轮作是该地区的主要耕作制度,该地区夏玉米出苗到拔节期一般在 7 月上旬至 7 月中旬、拔节到抽雄期在 7 月下旬至 8 月上旬、抽雄到乳熟期在 8 月中旬至 9 月上旬、乳熟到成熟期在 9 月中旬至 9 月下旬。通过王鹏新等[18]提出的基于时间序列叶面积指数傅里叶变换的作物种植区域提取方法提取了 2010—2018 年研究区域玉米种植区。

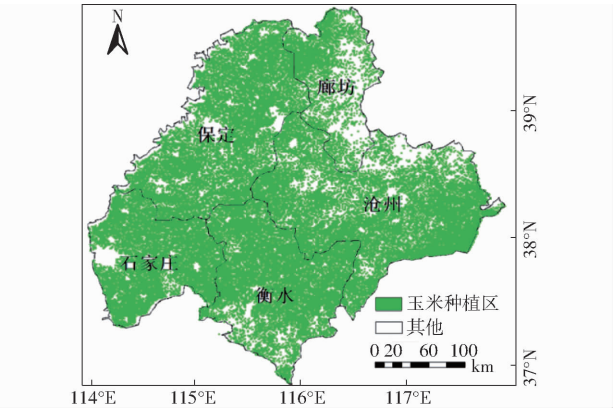

图 1　研究区域位置及玉米种植区(2010 年)

Fig. 1　Location of study area and planting area of maize(2010)

1.2　数据获取与处理

1.2.1　时间序列 VTCI 和 LAI 生成

选取 2010—2018 年每年 7—9 月 Aqua－MODIS 日地表温度产品 MYD11A1 及日地表反射率产品 MYD09GA,经 MRT 预处理后获得研究区域日 LST 和日 NDVI 产品,应用最大值合成技术生成每年 7—9 月旬时间尺度的 NDVI 和 LST 最大值合成产品,基于多年某一旬的 NDVI 和 LST 最大值合成产品,运用最大值合成技术分别生成多年的旬 NDVI 和 LST 最大值合成产品,基于每年 7—9 月旬 LST 最大值合成产品,运用最小值合成技术生成多年的旬 LST 最大-最小值合成产品。VTCI 取值范围为 0～1,其值越接近 0,表明越干旱,作物受水分胁迫程度越重,其值越接近 1,表明越湿润,作物受水分胁迫程度越轻或不受水分胁迫,VTCI 计算公式为

$$VTCI = \frac{L_{\max}(N_i) - L(N_i)}{L_{\max}(N_i) - L_{\min}(N_i)} \tag{1}$$

其中

$$L_{\max}(N_i) = a + bN_i \tag{2}$$

$$L_{\min}(N_i) = a' + b'N_i \tag{3}$$

式中　$L(N_i)$——在研究区域内,某一像素的 NDVI 值为 N_i 时的地表温度

$L_{\max}(N_i)$、$L_{\min}(N_i)$——研究区域当 NDVI 值为 N_i 时所有像素地表温度最大值和最小值

a、b、a'、b'——待定系数,由研究区域 LST 和 NDVI 的散点图近似得到

选取研究区域 2010—2018 年每年 7—9 月 MODIS 叶面积指数产品 MCD15A3H,该产品是基于 Terra 和 Aqua 卫星上的 MODIS 传感器获得的,与 MOD15A2 和 MYD15A2 产品相比,MCD15A3H 产品既有较高的时间分辨率(4 d)又有较高的空间分辨率(500 m),有利于作物长势监测及产量估测。利用 MRT 对产品进行预处理得到研究区域叶面积指数产品,原始叶面积指数产品由于云和大气等因素的影响存在数据骤降的现象,因此通过上包络线 S－G(Savitzky－Golay)滤波对原始叶面积指数产品进行平滑处理[18],经上包络线 S－G 滤波平滑处理后的叶面积指数更加符合玉米生长情况。为使 LAI 与 VTCI 具有相同的时间尺度,将玉米各旬所包含的多时相 LAI 的最大值作为各旬的 LAI 值,并对上包络线 S－G 滤波后的 LAI 进行归一化处理,最大值为 7,最小值为 0。

1.2.2　VTCI 和 LAI 计算

依据玉米 4 个主要生育时期的划分,将玉米各生育时期包含的多旬 VTCI 和 LAI 的平均值作为该生育时期的 VTCI 和 LAI 值,如将 7 月上旬至 7 月中旬 VTCI 的平均值作为出苗到拔节期的 VTCI 值。再叠加研究区域行政边界图,将各县(区)包含的所有像素的 VTCI 和 LAI 的平均值作为该县(区)的 VTCI 和 LAI 值。以此类推,计算得到研究区域 2010—2018 年各县(区)玉米各生育时期的 VTCI 和 LAI 值。

1.2.3　玉米单产数据的来源及异常数据处理

通过查阅《河北农村统计年鉴》得到研究区域各县(区)2010—2016 年玉米播种面积和总产量数据,玉米单产由总产量和播种面积计算得到。

将 VTCI 和 LAI 与玉米单产进行回归分析的残差的置信区间在[－4 000,4 000] kg/hm^2 以外的单产数据视为异常数据,在构建模型时将其剔除。

1.3　研究方法

随机森林回归对噪声数据集容忍度较高,对高维数据集具有良好的预测能力[19－20]。它是由一组没有联系的回归决策树$\{h(x,\theta_k), k=1,2,\cdots,K\}$构成的 K 棵集成决策树,表示为

$$h(x)=\frac{1}{N}\sum_{k=1}^{K}h(x,\boldsymbol{\theta}_k) \tag{4}$$

式中 x——各县(区)玉米 4 个生育时期 VTCI 或 LAI 值及玉米单产数据

K——决策树的数量

$\boldsymbol{\theta}_k$——独立同分布随机向量

为了提高模型的预测精度并防止出现过拟合情况,以随机森林回归算法结合袋装法得到训练样本子集,并结合随机子空间法得到节点分裂特征[21]。

(1)袋装法通过有放回地随机抽样,从原始样本数据集中重复抽样得到 K 个与原始样本数据集相等的训练样本 N,每个训练样本构成一棵决策树。每次进行 Bootstrap 重抽样时,未被抽中的样本的概率为 $(1-1/N)^N$,当 N 趋向于无穷大时,未被抽中样本的概率越接近 1/e,约为 0.368,即原始样本中有 36.8% 的数据未被抽中,这些数据被称为袋外数据(Out of bag, OOB),因其未参与回归树的构建,故可用来估计预测袋外数据误差(OOB 误差)及评估自变量对因变量的影响程度。另外,基于 OOB 预测误差可以检验模型的泛化能力,不需再使用测试集检验模型的精度。通过袋装法得到的 K 个训练样本都不相同,保证了回归树的差异性。

(2)随机子空间法通过袋装法得到 K 棵回归树后,每个分裂节点随机抽取所有变量(特征)中的 M_{try} 个变量(特征)作为当前节点分裂的特征子集,根据分类回归树(Classification and regression tree, CART)方法在特征子集中选择最优分裂方式进行分裂。通过随机子空间法得到的回归树具有随机性和独立性。在随机森林回归中,树的数量 K 和随机选择的节点分裂变量(特征)M_{try} 决定着模型的预测能力。

基于随机森林回归估测玉米单产的流程如图 2 所示。

(1)将研究区域各县(区)2010—2016 年玉米 4 个生育时期的 VTCI 或 LAI 值及玉米单产数据作为原始样本(共 357 组数据)输入模型,通过 Bootstrap 重抽样得到 K 个训练样本子集并生成 K 棵回归树。VTCI 和 LAI 估测玉米单产的 OOB 误差随树的数量 K 变化曲线如图 3 所示,可以看出,当 K 为 500 时,OOB 误差趋于平稳,故将 K 设为 500。

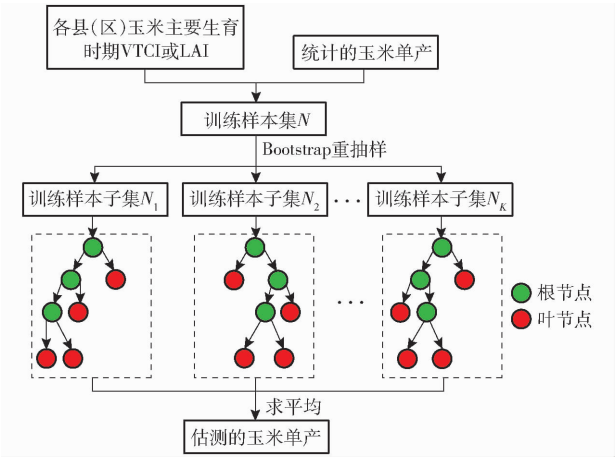

图 2 基于随机森林回归估测玉米单产的流程图

Fig. 2 Flow chart for estimating maize yield based on random forest regression

(2)在回归树生长过程中,每个节点进行分裂时,利用 CART 方法随机选取 M_{try} 个变量(随机森林回归算法一般选取总变量的 $\frac{1}{3}$ 个变量,因此 M_{try} 为 1)进行分裂,且分裂过程中不需要剪枝。

(3)每棵回归树自上向下分裂生长,直到到达某个叶子节点输出估测值,所有回归树构成随机森林。将所有回归树输出的玉米单产求平均值即可得到最终的玉米单产估测结果。

随机森林回归模型不但能精确地估测玉米单产,而且还可给出各个变量的重要性评分,即玉米 4 个生育时期 VTCI 或 LAI 对玉米单产的影响程度。基于基尼系数和基于 OOB 误差是常用的变量重要性评分的统计量,本研究中基于 OOB 估测误差得到各变量的重要性。若 $x_j(j=1,2,3,4)$ 为输入变量,则在第 k 棵树上的重要性 I_k 为随机置换变量前后袋

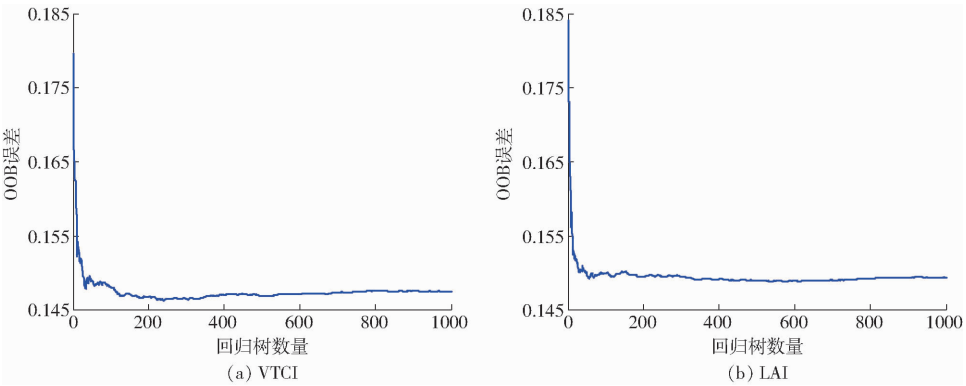

图 3 OOB 误差随回归树数量的变化曲线

Fig. 3 Changing curves of OOB errors with number of regression trees

外数据估测误差的差值[22]。其计算公式为

$$I_k(x_j) = \left[\sum_{n=1}^{N_{OOB}} I(f(x_n) = f_k(x_n)) - \sum_{n=1}^{N_{OOB}} I(f(x_n) = f_k(x'_n)) \right] \Big/ N_{OOB} \quad (5)$$

变量 x_j 在整个随机森林中的重要性得分为

$$I(x_j) = \sum_{k=1}^{K} I_k(x_j)/K \quad (6)$$

式中　N_{OOB}——袋外数据样本数

$f(x_n)$——袋外数据中第 n 个样本值

$f_k(x_n)$、$f_k(x'_n)$——随机置换变量前后第 k 棵树上的袋外数据第 n 个样本的估测值

$I(\cdot)$——判别函数，当 $f(x_n) = f_k(x_n)$ 或 $f(x_n) = f_k(x'_n)$ 时，取值为 1，否则为 0

由于随机性的引入，模型每次给出的变量重要性评分略有差异，故将 10 次运行结果的平均值进行归一化处理，作为各个变量的权重。

1.4　估产模型构建

通过随机森林回归方法确定玉米主要生育时期 VTCI 和 LAI 的权重，计算 2010—2018 年各县（区）加权 VTCI 和 LAI。对 2010—2016 年（除 2012 年，用来进行精度验证）加权 VTCI 和 LAI 与玉米单产进行回归分析，选取拟合程度最优的回归模型对 2012 年各县（区）的玉米单产进行估测及精度验证，并基于该模型逐像素估测 2010—2018 年研究区域的玉米单产。

2　结果与分析

2.1　玉米各生育时期的权重

基于随机森林回归模型运行 10 次输出的各变量重要性的平均值进行归一化处理，得到玉米各生育时期 VTCI 和 LAI 的权重（表 1）。可以看出，玉米拔节—抽雄期和抽雄—乳熟期的 VTCI 权重较大，说明受水分胁迫时对玉米单产的影响程度相对较大，主要是因为这两个时期对水分胁迫较敏感，抽雄期前后发生水分胁迫会导致幼穗发育不良，果穗偏小，雄穗在抽出 2 ~ 3 d 后失去散粉能力，甚至有的雄穗不能抽出，或抽穗时间延迟，导致秃尖增长，造成不同程度的玉米产量下降，水分胁迫较重的会造成雌穗部分不育甚至空秆。苗期—拔节期和乳熟—成熟期的 VTCI 权重相对较小，说明发生水分胁迫对玉米单产的影响较小，主要是苗期发生一定程度的水分胁迫会使根向下生长，有利于玉米植株后期的生长发育，且后期有充足水分时能够弥补之前减少的生长量，乳熟期之后穗粒已经形成，受水分影响不大[23]。LAI 对玉米单产的影响以抽雄—乳熟期和乳熟—成熟期较大，苗期—拔节期和拔节—抽雄期较小，表明生长前期 LAI 与玉米产量的相关性不大，主要是因为光合作用的产物用来进行以根系和叶片为中心的营养生长，抽雄期时 LAI 达到最大，玉米进入以果穗为中心的生殖生长阶段，LAI 与产量的相关性开始增大，这与姚小英等[24]的研究结果较一致。

表 1　玉米各生育时期的权重结果

Tab. 1　Weight results of each growth stage of maize

序号	VTCI				LAI			
	苗期—拔节期	拔节—抽雄期	抽雄—乳熟期	乳熟—成熟期	苗期—拔节期	拔节—抽雄期	抽雄—乳熟期	乳熟—成熟期
1	1.17	1.62	2.21	1.42	0.54	0.39	0.93	1.35
2	1.20	1.59	2.22	1.34	0.55	0.35	0.83	1.31
3	1.13	1.57	1.97	1.34	0.49	0.35	0.83	1.25
4	1.17	1.60	2.23	1.36	0.57	0.43	0.86	1.36
5	1.23	1.63	2.14	1.51	0.57	0.35	0.85	1.40
6	1.19	1.61	2.25	1.38	0.54	0.40	0.84	1.27
7	1.20	1.64	2.10	1.48	0.55	0.34	0.82	1.27
8	1.11	1.71	2.06	1.48	0.51	0.37	0.82	1.33
9	1.19	1.65	2.11	1.48	0.64	0.40	0.90	1.37
10	1.26	1.59	2.27	1.34	0.61	0.40	0.83	1.31
平均值	1.19	1.62	2.16	1.41	0.56	0.38	0.85	1.32
权重	0.19	0.25	0.34	0.22	0.18	0.12	0.27	0.43

2.2　单产估测模型选择

将随机森林回归方法计算得到的 2010—2016 年（除 2012 年）加权 VTCI 和 LAI 与玉米单产基于县域尺度进行线性回归分析，建立不同变量的单产估测模型（表 2）。结果表明，基于随机森林回归的加权 VTCI 和玉米单产的相关性最低（$R^2 = 0.001$），且没有通过显著性检验；加权 LAI 与玉米单产的相关性次之（$R^2 = 0.296$）；加权 VTCI 和 LAI 与玉米单产的相关性最高（$R^2 = 0.303$），模型达极显著水平（$P < 0.001$），表明 VTCI 和 LAI 与玉米单产呈显著

的正相关关系。因此,基于双变量估产模型的精度高于单变量模型的精度。基于随机森林回归双变量估产模型估测玉米单产时,玉米单产受 LAI 影响较大,VTCI 影响较小,原因可能是研究区域受人为因素的影响较大,当发生水分胁迫时,通过及时灌溉缓解了当地旱情,致使玉米单产对 VTCI 不敏感。综上所述,基于随机森林回归的双变量估产模型精度最高,可用于估测研究区域 2012 年各县(区)的玉米单产。

表 2 加权 VTCI 和 LAI 与玉米单产间的线性回归分析

Tab. 2 Linear regression analysis between weighted VTCI and LAI and maize yields

变量	估产模型	R^2	P
VTCI	$Y = -300VTCI + 6\,653$	0.001	0.700
LAI	$Y = 8\,431LAI + 2\,621$	0.296	<0.001
LAI、VTCI	$Y = 8\,571LAI - 1\,114VTCI + 3\,257$	0.303	<0.001

2.3 2012 年各县(区)玉米单产估测的精度评价

基于随机森林回归双变量估产模型及 2012 年加权 VTCI 和 LAI 对各县(区)玉米单产进行估测(表 3)。玉米估测单产与实际单产的相对误差以清苑区最低,为 0.35%,以海兴县最高,为 37.10%。其中,31 个县(区)玉米估测单产与实际单产的相对误差在 10% 以下,7 个县(区)在 10% ~15%,15 个县(区)在 15% 以上,53 个县(区)的平均相对误差为 9.85%,均方根误差为 824.77 kg/hm^2。个别县(区)如海兴县、盐山县的相对误差较大,原因可能是海兴县、盐山县濒临渤海,土壤盐渍化严重,农业生产条件较差,农田水利设施建设和机械化水平较低,不适宜种植经济作物,种植冬小麦和夏玉米是仅有的选择。近年来当地已采取改造重盐碱地的相关措施使玉米单产有所提高,但是玉米生产仍处于较低水平,玉米单产被高估,从而使估测单产与实际单产的相对误差较大。个别县(区)如正定县、藁城区和新乐市实际玉米单产较高,估测单产偏低,玉米单产被低估,原因可能是这几个县(区)是国家粮食丰产科技工程河北省项目区的核心区,田间管理及时,玉米单产受人为因素影响较大。

表 3 2012 年各县(区)玉米估测单产

Tab. 3 Estimated yields of maize in each county (district) in 2012

县(区)	实际单产/(kg·hm^{-2})	估测单产/(kg·hm^{-2})	相对误差/%	县(区)	实际单产/(kg·hm^{-2})	估测单产/(kg·hm^{-2})	相对误差/%
涿州市	6 163	6 352	3.07	肃宁县	6 625	7 066	6.66
高碑店市	7 949	7 160	-9.93	献县	5 823	6 514	11.87
雄县	6 415	6 663	3.86	沧县	5 019	5 348	6.55
定兴县	8 456	7 107	-15.96	黄骅市	5 182	4 837	-6.65
容城县	7 803	7 021	-10.02	泊头市	6 625	6 830	3.10
安新县	5 887	6 599	12.10	南皮县	6 398	6 498	1.57
高阳县	6 931	6 427	-7.27	孟村县	5 461	5 398	-1.16
徐水区	7 344	7 218	-1.71	盐山县	4 189	4 959	18.38
清苑区	7 463	7 489	0.35	海兴县	3 288	4 508	37.10
蠡县	6 536	6 868	5.08	东光县	7 513	6 529	-13.10
望都县	8 742	7 701	-11.91	吴桥县	8 103	6 809	-15.97
定州市	8 142	7 656	-5.97	安平县	6 075	7 449	22.62
安国市	7 980	8 234	3.18	饶阳县	6 272	6 555	4.51
博野县	8 185	8 022	-1.99	深州市	6 870	7 388	7.55
新乐市	8 265	6 788	-17.87	武强县	6 270	7 485	19.37
正定县	8 565	6 792	-20.70	武邑县	6 199	7 593	22.48
藁城区	8 655	7 207	-16.73	冀州市	6 049	7 270	20.18
无极县	8 175	7 493	-8.34	阜城县	6 085	7 252	19.17
深泽县	8 460	7 586	-10.33	故城县	6 427	6 951	8.15
辛集市	8 085	7 412	-8.32	景县	6 870	7 344	6.89
晋州市	8 190	6 927	-15.42	枣强县	7 109	7 523	5.82
栾城区	8 250	7 219	-15.86	大城县	5 952	6 315	6.11
赵县	8 190	7 388	-9.79	永清县	6 523	5 950	-8.78
高邑县	8 250	7 006	-15.08	固安县	6 240	6 347	1.71
任丘市	6 621	6 694	1.10	霸州市	6 303	6 410	1.70
河间市	6 499	6 638	2.15	文安县	6 875	6 167	-10.30
青县	5 469	5 436	-0.60				

为了进一步验证随机森林回归双变量估产模型的精度，基于 2012 年各县（区）玉米实际单产与估测单产进行线性回归分析。结果表明，估测单产与实际单产间呈显著的正相关关系（$P<0.001$），R^2 达到 0.540；估测单产与实际单产的均方根误差为 631.64 kg/hm^2，进一步说明基于随机森林回归双变量估产模型的精度较高，可用于研究区域玉米单产估测。

2.4　玉米估测单产的时空变化规律

基于随机森林回归双变量估产模型逐像素估测 2010—2018 年研究区域玉米单产（图 4），并逐像素统计玉米估测单产。结果表明，2010、2012、2013 年玉米估测单产相差不大，西部地区（包括石家庄市和保定市）玉米估测单产在 6 600 kg/hm^2 左右，东部地区（包括沧州市）在 6 100 kg/hm^2 左右，南部地区（包括衡水市）在 6 800 kg/hm^2 左右，北部地区（包括廊坊市）在 6 200 kg/hm^2 左右；2011 年玉米估测单产略高于 2010 年；2014 年玉米估测单产略低于 2013 年；2015、2016、2017 年玉米估测单产略高于 2014 年，西部地区在 6 900 kg/hm^2 左右，东部地区在 6 100 kg/hm^2 左右，南部地区在 7 000 kg/hm^2 左右，北部地区在 6 100 kg/hm^2 左右，2018 年西部地区和南部地区玉米单产在 7 000 kg/hm^2 左右，东部地区和北部地区在 6 500 kg/hm^2 左右。以 2017 年为例，西部地区玉米估测单产为 6 868 kg/hm^2，东部地区为 6 051 kg/hm^2，南部地区为 6 833 kg/hm^2，北部地区为 6 045 kg/hm^2。研究年份间 2011 年玉米单产最高，2014 年玉米单产较低，原因可能是 2011 年降水量充沛，玉米单产高于常年，2014 年玉米生育期内发生阶段性干旱且局部地区旱情较重，玉米单产下降。

3　讨论

课题组在陕西关中平原的冬小麦干旱监测及单产估测中采用客观赋权法如熵值法确定 VTCI 的权重[25]，构建的加权 VTCI 和冬小麦单产的回归模型精度较高，但熵值法基于指标的差异程度确定指标权重，异常数据对权重影响较大，且可能使权重与实

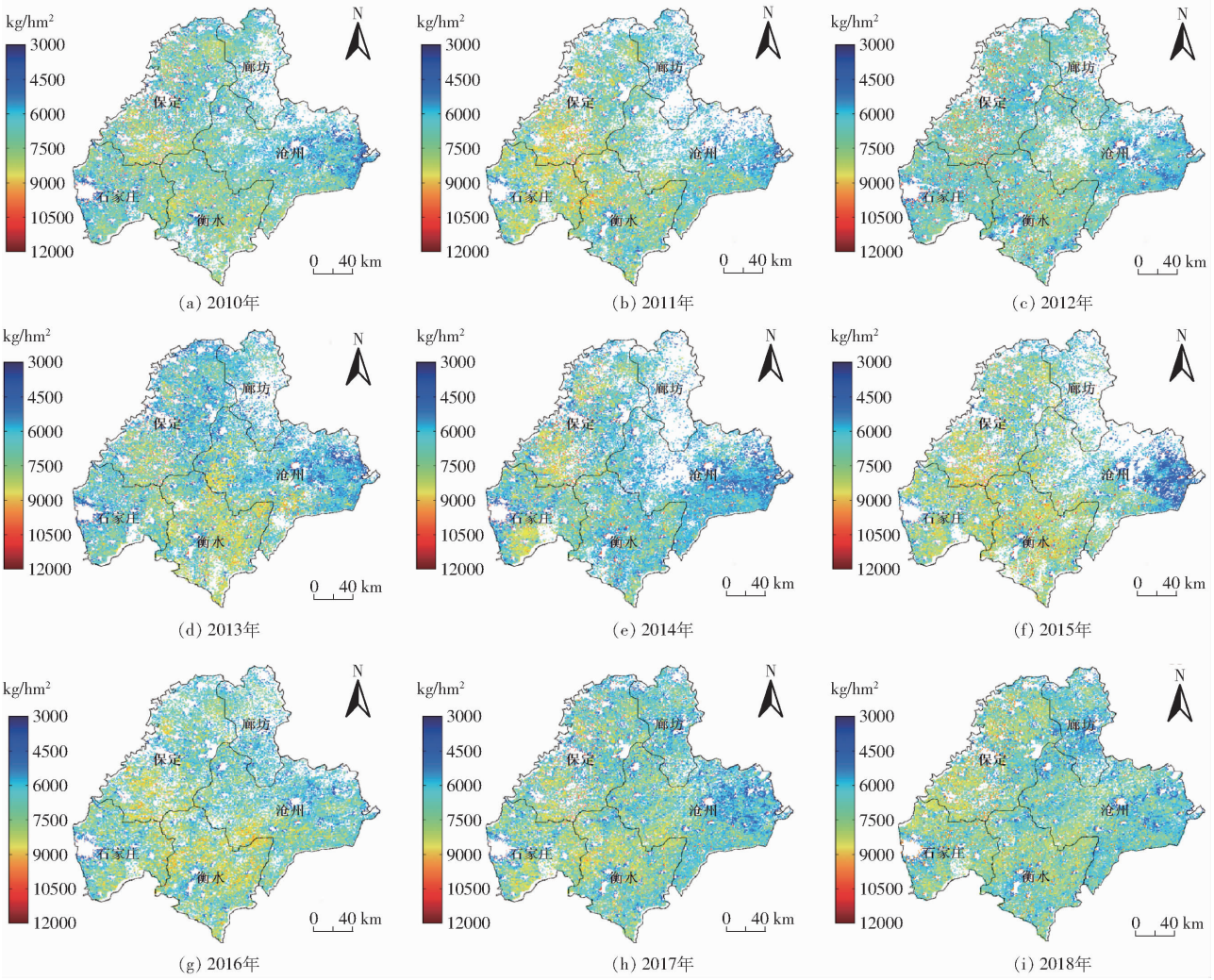

图 4　基于随机森林回归的玉米单产估测结果

Fig. 4　Estimate results of maize yields based on random forest regression

际相背,因此确定冬小麦主要生育时期 VTCI 的权重与实际水分胁迫对冬小麦单产的影响程度不符。在河北省中部平原地区应用随机森林回归确定玉米主要生育时期 VTCI 和 LAI 的权重,结果表明随机森林回归确定的 VTCI 权重以拔节—抽雄期、抽雄—乳熟期的权重较大,根据实际水分胁迫对玉米单产的影响程度[26]可以看出,基于随机森林回归的权重结果更加符合实际情况。主要因为干旱对玉米单产的影响具有非线性的特征,随机森林回归模型对于非平衡数据比较稳健,不易受到异常值的干扰,能有效处理非线性问题。虽然基于随机森林回归确定的玉米主要生育时期 VTCI 和 LAI 的权重较合理,但是未考虑农学先验知识,可通过结合主观赋权法如改进的层次分析法进一步修正随机森林回归得到的权重,使权重更加符合实际情况。另外水分胁迫也会影响玉米的生长状态,即 VTCI 和 LAI 之间可能存在多元共线性的问题,而随机森林回归模型对多元共线性不敏感,可以很好地预测多个变量的作用,因此随机森林回归模型的精度较高。

影响玉米单产的因素有很多,除了受到水分胁迫和生长状态的影响外,还受到其他因素如温度、洪涝灾害、田间管理、玉米品种等的影响。杨笛[27]通过模拟气候变化、肥料、种植面积、灌溉和品种 5 个驱动因子对黄淮海夏玉米区玉米单产的影响,表明肥料和品种在玉米增产中的作用和地位随时间在提高,种植面积的增长及灌溉系数的减少不利于玉米增产。通过查阅《河北农村统计年鉴》可以看出,研究年份间灌溉和肥料的使用较多,这可能与研究区域玉米高产有一定的联系。另外,个别年份发生灾害如 2016 年研究区域部分县(区)玉米苗期发生雹灾,影响玉米出苗,7 月又出现涝灾和病虫害使玉米单产略有下降。这些因素对玉米单产的影响不容忽视,综合考虑与玉米单产相关性较大的因素是今后研究的重点。

4 结论

(1)通过随机森林回归确定玉米主要生育时期 VTCI 和 LAI 的权重,构建加权 VTCI 和 LAI 与玉米单产的单变量和双变量估产模型。结果表明,基于随机森林回归的双变量估产模型精度最高。

(2)基于随机森林回归双变量估产模型估测 2010—2018 年研究区域玉米单产,结果表明,玉米估测单产在空间上的分布特征为西部地区最高、北部和南部次之、东部最低,年际间的分布特征为在波动中呈先减少后增加的趋势。

参 考 文 献

[1] MUELLER N D, GERBER J S, JOHNSTON M, et al. Closing yield gaps through nutrient and water management[J]. Nature, 2012, 490(7419): 254 – 257.

[2] 朱婉雪,李仕冀,张旭博,等. 基于无人机遥感植被指数优选的田块尺度冬小麦估产 [J]. 农业工程学报, 2018, 34(11): 78 – 86.
ZHU Wanxue, LI Shiji, ZHANG Xubo, et al. Estimation of winter wheat yield using optimal vegetation indices from unmanned aerial vehicle remote sensing [J]. Transactions of the CSAE, 2018, 34(11): 78 – 86. (in Chinese)

[3] 黄健熙,罗倩,刘晓暄,等. 基于时间序列 MODIS NDVI 的冬小麦产量预测方法 [J/OL]. 农业机械学报, 2016, 47(2): 295 – 301.
HUANG Jianxi, LUO Qian, LIU Xiaoxuan, et al. Winter wheat yield forecasting based on time series of MODIS NDVI [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(2): 295 – 301. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no=20160239&flag=1&journal_id=jcsam. DOI: 10.6041/j.issn.1000-1298.2016.02.039. (in Chinese)

[4] 贺振,贺俊平. 基于 NOAA – NDVI 的河南省冬小麦遥感估产[J]. 干旱区资源与环境, 2013, 27(5): 46 – 52.
HE Zhen, HE Junping. Estimation of winter wheat yield based on the NOAA – NDVI data [J]. Journal of Arid Land Resources and Environment, 2013, 27(5): 46 – 52. (in Chinese)

[5] 任建强,陈仲新,周清波,等. MODIS 植被指数的美国玉米单产遥感估测 [J]. 遥感学报, 2015, 19(4): 568 – 577.
REN Jianqiang, CHEN Zhongxin, ZHOU Qingbo, et al. MODIS vegetation index data used for estimating corn yield in USA [J]. Journal of Remote Sensing, 2015, 19(4): 568 – 577. (in Chinese)

[6] 王恺宁,王修信. 多植被指数组合的冬小麦遥感估产方法研究 [J]. 干旱区资源与环境, 2017, 31(7): 44 – 49.
WANG Kaining, WANG Xiuxin. Research on winter wheat yield estimation with the multiply remote sensing vegetation index combination [J]. Journal of Arid Land Resources and Environment, 2017, 31(7): 44 – 49. (in Chinese)

[7] LIAQAT M U, CHEEMA M J M, HUANG W, et al. Evaluation of MODIS and Landsat multiband vegetation indices used for wheat yield estimation in irrigated Indus Basin [J]. Computers and Electronics in Agriculture, 2017, 138: 39 – 47.

[8] 王蕾,王鹏新,李俐,等. 河北省中部平原玉米长势遥感综合监测 [J]. 资源科学, 2018, 40(10): 2099 – 2109.
WANG Lei, WANG Pengxin, LI Li, et al. Integrated maize growth monitoring based on gray correlation analysis and remote sense data in the central plain of Hebei Province [J]. Resources Science, 2018, 40(10): 2099 – 2109. (in Chinese)

[9] 王鹏新,龚健雅,李小文. 条件植被温度指数及其在干旱监测中的应用 [J]. 武汉大学学报(信息科学版), 2001, 26(5): 141 – 143.
WANG Pengxin, GONG Jianya, LI Xiaowen. Vegetation temperature condition index and its application for drought monitoring [J]. Geomatics and Information Science of Wuhan University, 2001, 26(5): 141 – 143. (in Chinese)

[10] 王蕾,王鹏新,李俐,等. 应用条件植被温度指数预测县域尺度小麦单产 [J]. 武汉大学学报(信息科学版),2018,43(10):1566 – 1573.
WANG Lei, WANG Pengxin, LI Li, et al. Wheat yield forecasting at county scale based on time series vegetation temperature condition index [J]. Geomatics and Information Science of Wuhan University, 2018, 43(10): 1566 – 1573. (in Chinese)

[11] 王鹏新,冯明悦,孙辉涛,等. 基于非线性特征的干旱影响评估研究[J/OL]. 农业机械学报,2016,47(10):325 – 331.
WANG Pengxin, FENG Mingyue, SUN Huitao, et al. Drought impact assessment based on nonlinear characteristics of drought [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(10): 325 – 331. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20161041&flag = 1&journal_id = jcsam. DOI: 10.6041/j.issn.1000-1298.2016.10.041. (in Chinese)

[12] 王鹏新,冯明悦,孙辉涛,等. 基于主成分分析和 Copula 函数的干旱影响评估研究[J/OL]. 农业机械学报,2016,47(9):334 – 340.
WANG Pengxin, FENG Mingyue, SUN Huitao, et al. Drought impact assessment based on principal component analysis and Copula function [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(9): 334 – 340. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20160945&flag = 1&journal_id = jcsam. DOI: 10.6041/j.issn.1000-1298.2016.09.045. (in Chinese)

[13] 林卉,梁亮,张连蓬,等. 基于支持向量机回归算法的小麦叶面积指数高光谱遥感反演[J]. 农业工程学报,2013,29(11):139 – 146.
LIN Hui, LIANG Liang, ZHANG Lianpeng, et al. Wheat leaf area index inversion with hyperspectral remote sensing based on support vector regression algorithm [J]. Transactions of the CSAE, 2013, 29(11): 139 – 146. (in Chinese)

[14] HOLZMAN M E, CARMONA F, RIVAS R, et al. Early assessment of crop yield from remotely sensed water stress and solar radiation data [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 145: 297 – 308.

[15] 王鹏新, 孙辉涛, 解毅,等. 基于 LAI 和 VTCI 及粒子滤波同化算法的冬小麦单产估测 [J/OL]. 农业机械学报, 2016, 47(4):248 – 256.
WANG Pengxin, SUN Huitao, XIE Yi, et al. Winter wheat yield estimation based on particle filter assimilation algorithm and remotely sensed LAI and VTCI [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(4): 248 – 256. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20160433&flag = 1&journal_id = jcsam. DOI: 10.6041/j.issn.1000-1298.2016.04.033. (in Chinese)

[16] 余坤勇,姚雄,邱祈荣,等. 基于随机森林模型的山体滑坡空间预测研究[J/OL]. 农业机械学报,2016,47(10):338 – 345.
YU Kunyong, YAO Xiong, QIU Qirong, et al. Landslide spatial prediction based on random forest model [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(10): 338 – 345. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20161043&flag = 1&journal_id = jcsam. DOI: 10.6041/j.issn.1000-1298.2016.10.043. (in Chinese)

[17] AHMAD M W, REYNOLDS J, REZGUI Y. Predictive modelling for solar thermal energy systems: a comparison of support vector regression, random forest, extra trees and regression trees [J]. Journal of Cleaner Production, 2018, 203: 810 – 821.

[18] 王鹏新,荀兰,李俐,等. 基于时间序列叶面积指数傅里叶变换的作物种植区域提取[J]. 农业工程学报, 2017,33(21):207 – 215.
WANG Pengxin, XUN Lan, LI Li, et al. Extraction of planting areas of main crops based on Fourier transformed characteristics of time series leaf area index products[J]. Transactions of the CSAE, 2017, 33(21): 207 – 215. (in Chinese)

[19] BREIMAN L. Random forest [J]. Machine Learning, 2001, 45(1): 5 – 32.

[20] LQBAL F, LUCIEER A, BARRY K. Poppy crop capsule volume estimation using UAS remote sensing and random forest regression [J]. International Journal of Applied Earth Observation and Geoinformation, 2018, 73: 362 – 373.

[21] 钟登华,田耕,关涛,等. 基于混沌时序-随机森林回归的堆石坝料加水量预测研究 [J]. 水力发电学报,2018,37(8):1 – 12.
ZHONG Denghua, TIAN Geng, GUAN Tao, et al. Prediction of rockfill dam material watering volume based on chaotic time series and random forest regression [J]. Journal of Hydroelectric Engineering, 2018, 37(8): 1 – 12. (in Chinese)

[22] GONG H R, SUN Y R, SHU X, et al. Use of random forests regression for predicting IRI of asphalt pavements [J]. Construction and Building Materials, 2018, 189: 890 – 897.

[23] 张淑杰,张玉书,纪瑞鹏,等. 水分胁迫对玉米生长发育及产量形成的影响研究 [J]. 中国农学通报,2011,27(12):68 – 72.
ZHANG Shujie, ZHANG Yushu, JI Ruipeng, et al. Influences of water stress on growth and development of maize and yield [J]. Chinese Agricultural Science Bulletin, 2011, 27(12): 68 – 72. (in Chinese)

[24] 姚小英, 李晓薇, 王禹锡, 等. 西北干旱区旱地玉米叶面积指数与气象因子及生物量的关系 [J]. 自然资源学报, 2012,27(11):1881 – 1889.
YAO Xiaoying, LI Xiaowei, WANG Yuxi, et al. Relationship among LAI and meteorological factors and biomass of maize in dry-farming areas of northwestern China [J]. Journal of Natural Resources, 2012, 27(11): 1881 – 1889. (in Chinese)

[25] 李艳,王鹏新,刘峻明,等. 基于条件植被温度指数的冬小麦主要生育时期干旱监测效果评价——I. 因子权重排序法和熵值法组合赋权 [J]. 干旱地区农业研究,2013,31(6):159 – 163.
LI Yan, WANG Pengxin, LIU Junming, et al. Application of temperature condition index to evaluate the drought monitoring effect in main growing period of winter wheat—I. Factor weight sorting method and entropy method [J]. Agricultural Research in the Arid Areas, 2013, 31(6): 159 – 163. (in Chinese)

[26] 白向历,孙世贤,杨国航,等. 不同生育时期水分胁迫对玉米产量及生长发育的影响 [J]. 玉米科学,2009,17(2):60 – 63.
BAI Xiangli, SUN Shixian, YANG Guohang, et al. Effects of water stress on maize yield at different growth stages [J]. Journal of Maize Sciences, 2009, 17(2): 60 – 63. (in Chinese)

[27] 杨笛. 中国玉米产量增长的驱动因素分析 [D]. 北京: 中国农业科学院,2017.
YANG Di. Contribution analysis of driving factors for maize yield growth in China [D]. Beijing: Chinese Academy of Agricultural Sciences, 2017. (in Chinese)