doi:10.11676/qxxb2020.007

随机森林机器算法在江苏省小麦赤霉病病穗率预测中的应用*

徐　敏[1]　徐经纬[2]　谢志清[1]　高　苹[1]　李亚春[1]　缪璟秋[3]

XU Min[1]　XU Jingwei[2]　XIE Zhiqing[1]　GAO Ping[1]　LI Yachun[1]　MIAO Jingqiu[3]

1. 江苏省气候中心，南京，210008

2. 南京信息工程大学气象灾害教育部重点实验室/气象灾害预报预警与评估协同创新中心，南京，210044

3. 苏州市吴中区东山气象站，苏州，215107

1. *Climate Center of Jiangsu Province*，*Nanjing* 210008，*China*

2. *Key Laboratory of Meteorological Disaster*，*Ministry of Education*（*KLME*）*/Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters*（*CIC-FEMD*）*/Nanjing University of Information Sciences & Technology*（*NUIST*），*Nanjing* 210044，*China*

3. *Dongshan Meteorological Station in Wuzhong*，*Suzhou* 215107,*China*

2019-07-08 收稿，2019-09-11 改回。

徐敏，徐经纬，谢志清，高苹，李亚春，缪璟秋. 2020. 随机森林机器算法在江苏省小麦赤霉病病穗率预测中的应用. 气象学报，78(1)：143-153

Xu Min，Xu Jingwei，Xie Zhiqing，Gao Ping，Li Yachun，Miao Jingqiu. 2020. Application of the random forest machine algorithm in forecasting diseased panicle rate of wheat scab in Jiangsu province. ***Acta Meteorologica Sinica*****，78(1):143-153**

Abstract　The identification of meteorological and biotic factors that have significant impacts on wheat scab and the development of models for predicting diseased panicle rates at different stages are of remarkable significance for improving the ability to predict scab seriousness and protecting ecological environment of farmlands. On the basis of observations of diseased panicle rate and winter wheat phenology as well as daily meteorological elements in 13 cities in Jiangsu Province of China during the period from 2002 to 2008, the dominant meteorological elements that affect diseased panicle rate are identified, and the contributions of individual elements to diseased panicle rate are assessed for different phonological stages in various regions. Models that are initialized at different times for predicting diseased panicle rates are developed using the random forest（RF）regression algorithm. The reliability of the models is verified against observations of diseased panicle rates. Meteorological and biotic factors during the heading and flowering stage have the largest contribution to final diseased panicle rates, followed by that in the jointing stage and overwintering period. The dominant factors that determine final diseased panicle rates are relative humidity, the total number of consecutive rainy days larger than 3 d, and sunshine during the heading and flowering stages. Sunshine duration, precipitation, relative humidity and rainy days during the jointing stage have significant influences on final diseased panicle rates. Temperature and snowfall during the overwintering period have large precursor impact on final diseased panicle rates. The identified relative importance of key variables in each growth period is consistent with the theory on the development, release, infection, and epidemic of scab. The accuracy of models predicting diseased panicle rates based on RF algorithm varies with the number of critical characteristic variables, regions, the value of parameter M_{try}, and the growth period. The earliest time when the models can be used to yield useable prediction of diseased panicle rates is the beginning of March. The longest valid forecast time of the models is about 3 months. With the time approaching the maturity period and increases in the number of important characteristic variables as inputs, the accuracy of the modes increases and the discrepancy between predicted and observed

* 资助课题：2020 年国内外作物产量气象预报专项、江苏省气象局科研基金（KM201906）和中国气象局气象关键技术集成重点项目（CMAGJ2015Z02）。

作者简介：徐敏，主要从事气象条件对病害的影响规律研究。E-mail：amin0506@163.com

diseased panicle rates is significantly reduced. Models have better skills in predicting medium and serious categories of scab. This study indicates that the RF algorithm is able to provide reliable prediction of scab and thus has a great application potential.

Key words Wheat scab, Random forest method, Forecast of the diseased panicle rate

摘 要 基于2002—2018年江苏省13个市的小麦赤霉病病穗率资料与生育期观测资料、相应时段内的逐日气象数据,应用随机森林机器学习算法,分生育期、分区域定量评估影响病穗率的主要气象因子特征变量和贡献率,按不同起报时间建立预测模型并进行验证。结果表明,各生育期重要特征变量贡献率的排序为:抽穗扬花期>拔节期>越冬期。抽穗扬花期湿度、连续≥3 d的雨日和日照对赤霉病起主导作用,拔节期日照、降雨量、湿度和雨日与越冬期气温和降雪对赤霉病均具有前期影响,甄别出的重要特征变量排序结果符合赤霉病菌发育、释放、侵染和流行规律;基于随机森林算法建立的病穗率预测模型的精度与重要特征变量个数、赤霉病发生区域、M_{try}参数设定、生育期有关;最早可在3月初进行预测,预测时效近3个月,起报时间越接近乳熟期,输入的重要特征变量越多,则病穗率预测准确率越高,病穗率模拟值与实测值的波动趋势完全一致,对赤霉病"中等"和"偏重"等级模拟效果好,表明随机森林算法在赤霉病预测中有较高的可靠性和业务应用潜力。

关键词 小麦赤霉病,随机森林法,病穗率预测

中图法分类号 P49

1 引 言

小麦在世界范围内种植面积广泛,既是人类主要的食物来源,又是重要的工业原料(王晓曦等,2008)。赤霉病是影响小麦产量和品质的主要病害,含有致呕毒素和类雌性毒素,容易导致人畜中毒。在中国、日本、东南亚、美国等主要粮食产区,赤霉病均呈增多趋势(陆维忠,2001)。20世纪90年代以来,中国赤霉病年发生面积均在4.15×10^6 hm^{2}以上,危害程度重,引起减产的幅度大(Starkey, et al, 2007),尤其是江淮麦区,春季雨量充沛,不同程度的赤霉病时有发生。2000年以来,受小麦—玉米、小麦—水稻轮作和秸秆还田等耕作制度变化和极端天气、气候事件的影响,中国小麦主产区赤霉病重发频率明显上升,2003、2012、2015、2016、2018年均为偏重以上流行,给小麦产量和品质造成了严重影响(曾娟等,2013;黄冲等,2019)。赤霉病属于典型的气象型病害,赤霉病菌生长、发育、繁殖、侵染和流行与温度、湿度、光照、风等气象要素密切相关(张汉琳,1987;吴春艳等,2003);同时,气象条件也影响小麦生长发育,进而影响赤霉病菌的易感生育期。大量研究(Champeil, et al, 2004;侯明生等,2006;肖晶晶等,2011)认为:小麦赤霉病流行程度主要取决于抽穗扬花期的温度和湿度的适宜情况,在满足一定的温度条件下,若阴雨天气持续时间长,发病就重,反之则发生程度较轻。随着气候变暖,抽穗扬花期累计雨日和相对湿度对赤霉病的影响权重增大,温度的影响权重相对减小(徐敏等,2019)。此外,赤霉病流行程度与抽穗前的降水量关系密切(姜明波等,2018),抽穗前若降水多,赤霉病菌子囊壳形成多,为赤霉病的流行创造了有利条件。赤霉病菌的流行规律大致可以分为三个阶段(贾金明,2002):秋末菌源体形成期(上年10月下旬—11月下旬),气象条件影响进入越冬期菌源量的多少;赤霉病菌越冬休止期(上年12月—当年2月或3月),气象条件影响赤霉病菌的越冬存活率;赤霉病菌发育成熟期(当年4月—5月中旬),气象条件直接影响赤霉病的发病与流行程度。由此可见,赤霉病菌生长到流行的整个过程都会受到气象因子的影响,不仅受同期气象因子影响,也受前期气象因子的影响,因子之间可能存在非线性的影响关系且过程复杂,传统统计方法在定量反映不同气象因子在小麦不同生育期对赤霉病发生、发展的影响程度存在困难,不能很好地描述诱发赤霉病形成的复杂过程,因此,寻找新的度量特征因子重要性的方法显得尤为重要。

近年来,随着人工智能技术的飞速发展,随机森林(Random Forest)等机器学习算法在特征变量重要性评估和预测模型构建等方面开始凸显优势。随机森林是Breiman(2001)将Bagging集成学习理论与随机子空间结合提出的一种组合分类智能算法。该算法能有效解决高维变量问题,可以评估变量的重要性,具备分析复杂相互作用分类特征的能力,训练速度快,收敛规则遵循大数定律、泛化误差具有收敛性,不易产生过拟合(Iverson, et al, 2008)。大量的理论和实证研究表明随机森林法是一种自然的非线性建模工具,是目前数据挖掘、生物信息学最热门

的前沿研究领域之一。在生态学、医学、管理学、经济学等众多领域得到了广泛应用，如作物分类（石礼娟等，2017；王利民等，2018）、灾害风险评估（吴孝情等，2017；赖成光等，2015）、生物物种分布影响因素评估（张雷等，2011a，2011b）等均取得了较好的效果，构建的非线性预测模型预测精度高。随机森林算法所具有的计算特性和优点理论上为评估多种气象因子对小麦赤霉病的影响重要性、全面理解不同生育期影响赤霉病发生流行的主导气象因子和非主导气象因子提供了一种新思路。

为此，本研究根据小麦赤霉病菌流行规律的3个阶段，选取相应时段内的温度、湿度、降雨或降雪、日照等多类别气象因子为特征变量，病穗率为目标变量，利用随机森林算法对关键生育期影响病穗率的特征变量进行重要性排序，筛选出重要特征变量，在此基础上，根据不同的起报时间建立阶段性预测模型，以期提升病穗率预测时效、提高服务能力。

2　资料与方法

2.1　研究区概况

近10年江苏省小麦种植面积维持在2.10×10^6 hm^2左右，占全国小麦面积的9%，位列全国第4，是最大的弱筋小麦主产省份，对中国小麦产业至关重要。江苏小麦赤霉病流行频率高，发生程度重，每年发生面积超过6.7×10^5 hm^2，是对小麦威胁最大的病害。江苏省赤霉病具有3个特征：（1）田间菌源充足，各地稻桩子囊壳丛带菌率均远超大发生指标，且有逐年增加的趋势，沿江、里下河、沿淮部分地区高于10%；（2）自然发病程度重，近年来小麦赤霉病每年均有自然发病重的田块，重发田块病穗率在20%以上，重发年份个别失治田块甚至造成绝收；（3）品种间发病程度差异大，江苏省各地主栽品种普遍易感赤霉病或耐病性较弱，特别是沿海、沿淮、里下河北部高感品种种植面积较大。

2.2　数据资料

（1）小麦赤霉病病穗率数据：收集整理2002—2018年苏州、无锡、常州、南通、镇江、南京、扬州、泰州、盐城、淮安、宿迁、徐州、连云港13个市的病穗率数据，所谓“病穗率”是指发病的小麦穗数占调查总穗数的比率。该数据由江苏省农业植保部门在不进行人为化学防治的麦田，按照国家标准GB/T 15796-2011《小麦赤霉病测报技术规范》观测，在5月末计算出病穗率。调查时间是从抽穗始期开始，每日观察，始见病穗后，每3 d调查一次，至病情稳定为止；调查地点是选择当地一块系统调查田，面积不小于6.67×10^{-2} hm^2，栽种当地代表性品种2—3个，其中必须有一个感病品种，分早、中、迟3个播期，播期间隔10—15 d，每个品种种植面积不小于6.7×10^{-3} hm^2，生长期均不喷杀菌剂防治；调查方法是在已发现病穗的田块随机固定500穗，然后调查病穗数。

（2）小麦生育期观测资料：来自2002—2018年江苏省气象局10个农业气象观测站的《作物生长发育状况记录年报表》，生育期由专业的农业气象技术人员按照《农业气象观测规范 冬小麦》（QX/T 200-2015）观测所得，观测站点分别为昆山、沭阳、大丰、如皋、兴化、淮安、盱眙、滨海、赣榆、徐州。

（3）小麦生育期内气象资料：来自江苏省气象局2002—2018年江苏13个市逐日气温、降水量、相对湿度、日照时数、天气现象、风速等。

2.3　资料预处理

按照江苏省冬小麦的生育期，结合赤霉病菌流行规律，将分析时段分为：越冬期（上年12月—当年2月，即冬季）、拔节期（3月）、抽穗扬花前期（4月上旬）、抽穗扬花期。在建病穗率预测模型前，初步选出对病穗率有影响且符合生物学意义的气象因子非常关键，主要依据已有研究结果进行初步筛选。抽穗扬花期是赤霉病菌侵染关键期，主要影响因子是温度、湿度、降水、光照和风，其中温、湿度的匹配程度是关键（张汉琳，1987；肖晶晶等，2011；徐敏等，2019）。抽穗扬花前期，赤霉病菌主要影响因子是温度、湿度、降水、光照（吴春艳等，2003）。拔节期和越冬期对赤霉病具有前期影响，其中拔节期影响赤霉病菌的主要气象因子是气温、降水、湿度、光照（贾金明，2002；刁春友等，2006；姜明波等，2018）；越冬期影响赤霉病菌的主要气象因子是气温，霍治国等（2009）指出，冬季高温易使小麦赤霉病发病时的菌源数增多，一定程度上会增加小麦感染病菌的风险程度。另外，考虑到冬季降雪会对气温产生影响，尤其积雪融化造成的低温不利于赤霉病菌存活，所以越冬期影响因子还加入了积雪深度。最终初步入选的气象因子见表1。

表 1 用于分析评估对小麦赤霉病病穗率影响重要性的气象因子

Table 1 Meteorological factors used for analysis and evaluation of their importance on influencing the diseased panicle rate of wheat scab

时段	气象因子	个数
越冬期(上年12月—当年2月)	冬季平均最高气温、冬季最低气温≤0℃日数、冬季累计最大积雪深度、冬季雪深≥1 cm日数、冬季雪深≥5 cm日数、冬季雪深≥10 cm日数、冬季雪深≥20 cm日数、冬季雪深≥30 cm日数、冬季降水量、冬季日降水量≥0.1 mm日数、冬季日照时数、冬季平均相对湿度	12
拔节期(3月)	3月平均气温、3月最低气温≤0℃日数、3月降水量、3月降水量≥0.1 mm日数、3月降水量≥1 mm日数、3月降水量≥5 mm日数、3月降水量≥10 mm日数、3月降水量≥25 mm日数、3月平均相对湿度、3月日照时数	10
抽穗扬花前期(4月上旬)	4月上旬平均气温、4月上旬降水量、4月上旬降水量≥0.1 mm日数、4月上旬平均相对湿度、4月上旬累计日照时数	5
抽穗扬花期(4月中旬—5月上旬)	抽穗扬花期平均气温、抽穗扬花期累计降水量、抽穗扬花期累计雨日、抽穗扬花期降水连续≥3 d的雨日总数、抽穗扬花期平均相对湿度、抽穗扬花期平均日日照时数、抽穗扬花期平均风速、抽穗扬花期平均最大风速	8

由于江苏省南北跨度大,不同区域的抽穗扬花期存在一定差异,按照气候相似性原则,综合考虑农业区划,江苏省可分为 3 个区:苏南(苏州、无锡、常州、南京、镇江)、苏中(南通、扬州、泰州、淮安、盐城)、苏北(宿迁、徐州、连云港)。通过历年生育期观测资料,计算出苏南、苏中、苏北抽穗扬花期平均起止时间分别为:4 月上旬—中旬、4 月中旬—下旬、4 月下旬—5 月上旬。

2.4 随机森林算法基本原理

随机森林算法是以决策树为基分类器的一个集成学习模型$\{h(X,\theta_k);k=1,2,\cdots L\}$,$\{\theta_k\}$表示独立同分布的随机变量,输入特征变量$X$时,每一棵树只投一票给其认为最佳的分类结果。所谓决策树(Han, et al, 2007)是单个分类器,是一种从无次序、无规则的训练样本中推理出决策树表示形式的分类规则的方法,相当于一种布尔函数。随机森林的分类结果由每棵树投票中得票数最多的类确定(Biau, 2012),最终分类决策见式(1)

$$H(x)=\arg\max_{Y}\sum_{i}I(h_i(x)=Y) \qquad (1)$$

式中,$H(x)$表示随机森林模型,$h_i(x)$表示每个决策树分类器,Y为目标变量,即病穗率,$I(h_i(x)=Y)$为指示性函数。

随机森林算法是高维学数据分析方法之一,主要用于高维数据分类和回归,并可计算出自变量对因变量的重要性评分(Donnelly, et al, 1996)。该算法采用的是自助抽样方法,运算过程中涉及决策树棵数(N_{tree})和节点数(M_{try})两个参数的设定。一般而言,模型的计算量与每次生成的树的数量成正比,N_{tree}增加时,在模型预测精度不能提高的情况下,N_{tree}值设定应尽可能小。M_{try}值要在模型构建过程中通过逐次计算来挑选最优值,回归模型中一般为变量个数的三分之一。由于随机森林算法对样本数据的量纲和单位不敏感,所以运算时无需对样本数据进行归一化处理。

2.5 优选重要特征变量的方法

随机森林算法是通过预测精度法计算每个特征变量的重要性,利用该算法本身所具有的变量重要性度量可以对特征变量的重要性进行排序,然后从中筛选出对最终结果影响较大的特征变量,删除一些和目标变量无关或者冗余的特征变量,即选出重要性靠前的特征。从而简化特征数据集,使得预测模型更精确。在随机森林模型中评价特征变量重要性的主要指标是精度平均减少值(I_{MSE})和节点不纯度减少值(I_{NP})。I_{MSE}是指变量随机取值后模型估算误差相对于原来误差的升高幅度,I_{NP}是指变量对各个决策树节点的影响程度,I_{MSE}或I_{NP}值越大,说明该变量越重要,反之则相对不重要。文中采用I_{MSE}作为变量重要性的评价指标(王超等, 2019)。

3 结果与分析

3.1 分生育期、分区域重要特征变量的筛选与评价

针对苏南、苏中、苏北 3 个区域,按照越冬期、拔节期、抽穗扬花前期、抽穗扬花期 4 个时段,通过随机森林算法,以各生育期气象因子(表 1)为输入向量(2002—2018 年 13 个市,每年 4 个时段共 35 个气象因子,累计 7735 个气象因子样本),以病穗率为输出向量(2002—2018 年 13 个市,共 221 个病穗率样

本），分区域分生育期对输入向量进行重要性排序，计算各特征向量的 I_{MSE}。

在成百上千次的机器学习过程中，并非每一次计算出的变量重要性排序结果都完全一致（Verikas, et al, 2011），此时可通过计算各区域各生育期 50 次模拟结果的 I_{MSE} 平均值来进行重要性排序，筛选出重要特征变量再进行随机森林建模可降低不重要变量对模型精度的干扰。以苏南为例，越冬期（图 1a）排在前 4 位的特征变量依次是：冬季平均最高气温、冬季雪深≥1 cm 日数、冬季日降水量≥0.1 mm 日数、冬季累计最大积雪深度。冬季累计最大积雪深度与冬季雪深≥5 cm 日数存在较为明显的拐点，将出现拐点前的特征变量确定为相对重要的变量（王超等，2019），则认为冬季平均最高气温对病穗率的重要性大于其他同期变量。从植物病害生理学角度，在越冬期，气温偏高或降水日数多则利于赤霉病菌的存活，冬季雪深≥1 cm 的日数越多或冬季累计最大积雪深度越大则不利于赤霉病菌的存活。从冬季平均最高气温、冬季雪深≥1 cm 日数、冬季降水量≥0.1 mm 日数、冬季累计最大积雪深度与病穗率的相关性也能反映这一关系，这 4 个特征变量与病穗率的相关系数分别为 0.132、−0.172、0.150、−0.124，均通过了 0.05 的显著性 t 检验。拔节期（图 1b）相对重要的特征变量依次为：日照时数、平均相对湿度、累计降水量、最低气温≤0℃ 日数、降水量≥1 mm 日数，这 5 个特征变量与病穗率的相关系数分别为：−0.403、0.460、0.442、0.229、0.304，均通过了 0.001 的显著性 t 检验。拔节期是小麦越冬后的关键生育期，决定着成穗率的高低，若日照偏少、降雨频繁且雨量偏多、田间湿度持续偏大，即若阴雨寡照的天气偏多，则会影响植株营养体的增大，生长缓慢，易感染赤霉病菌；若天气回暖后气温急剧下降，最低气温＜0℃ 时会发生冻害，所以当拔节期＜0℃ 的天数偏多时，也容易影响植株体的生长，存在后期感染赤霉病菌的风险。抽穗扬花期（图 1c）相对重要的特征变量依次为：抽穗扬花期平均相对湿度、抽穗扬花期降水连续≥3 d 的雨日总数、抽穗扬花期累计降水量、抽穗扬花期平均日日照时数、抽穗扬花期累计雨日、抽穗扬花期平均气温，该生育期是赤霉病菌侵染的关键期，若降雨偏多，尤其当持续降雨≥3 d 的雨日总数多，导致田间相对湿度高，加上气温高，则非常有利于病菌孢子释放、侵染、流行，前 5 个特征变量与病穗率的相关系数分别为：0.428、0.338、0.286、−0.290、0.278，均通过了 0.001 的显著性 t 检验。平均相对湿度、持续降雨≥3 d 的雨日数、累计降水量、累计雨日与病穗率呈显著正相关、累计日照与

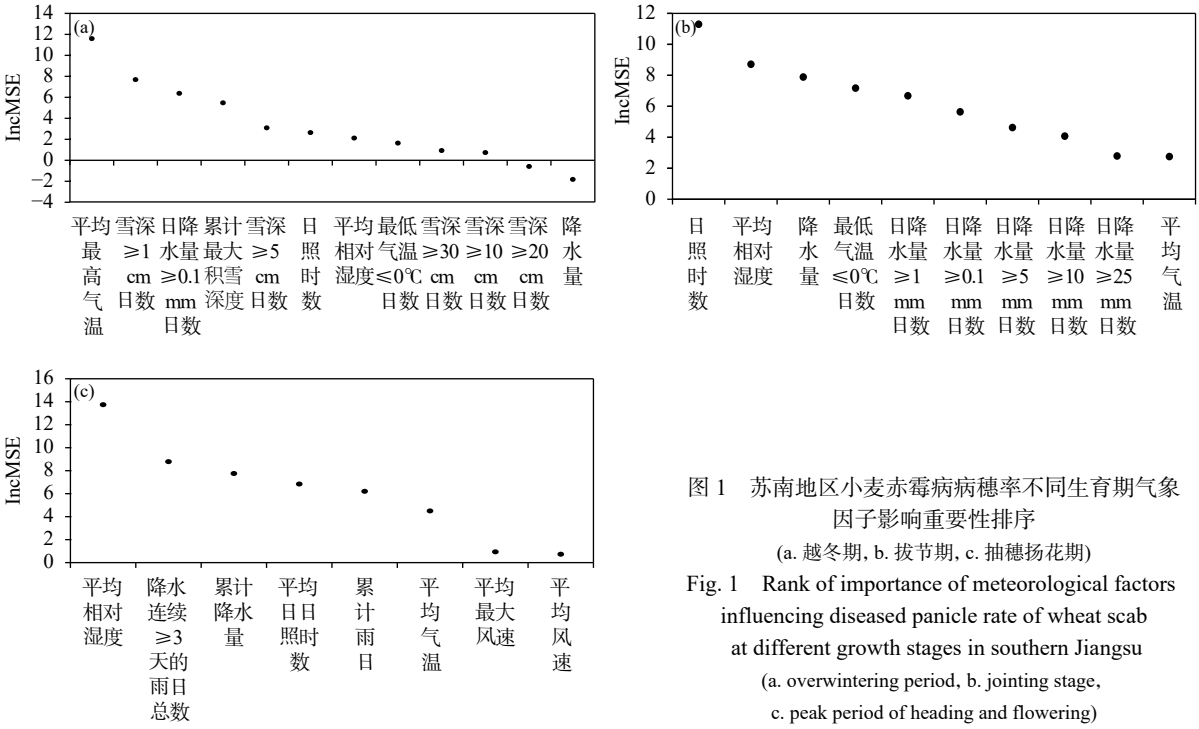

图 1　苏南地区小麦赤霉病病穗率不同生育期气象因子影响重要性排序

(a. 越冬期, b. 拔节期, c. 抽穗扬花期)

Fig. 1　Rank of importance of meteorological factors influencing diseased panicle rate of wheat scab at different growth stages in southern Jiangsu

(a. overwintering period, b. jointing stage, c. peak period of heading and flowering)

病穗率呈显著反相关,因为日照多意味着天气晴好,对赤霉病的发生、发展具有抑制作用,抽穗扬花期平均气温与病穗率的相关性不明显,主要是因为近年来气温条件通常都满足赤霉病发生、发展的要求。

按照苏南地区的计算思路,对苏中、苏北各生育期影响病穗率的重要特征变量也进行了筛选(表 2)。不同区域间由于赤霉病发生概率的差异以及地理气候等不同,筛选出的重要特征变量也存在一定差异,但对病穗率具有主导作用的变量基本一致。

3.2 按不同起报时间建立病穗率预测模型的思路和模型准确率

赤霉病可危害麦类的幼苗、茎秆和麦穗,苗期危害形成苗腐,拔节期形成茎秆基腐,其中以危害麦穗的损失最大,即赤霉病对不同生育阶段的小麦具有不同的影响,且该影响是个连续过程,这一规律为分时段建立病穗率预测模型提供了可行性。分时段病穗率预测模型的构建思路是:以表 2 中筛选出的影响病穗率的重要特征变量为输入向量,若起报时间是 3 月初,则选用越冬期的重要特征变量为预报因子;若起报时间是 4 月初,则选用越冬期和拔节期的重要特征变量为预报因子,随着生育进程不断推进,预报因子在逐步增多,树节点预选的变量个数 M_{try} 根据预报因子总数而定,决策树棵数 N_{tree} 设定为 600,病穗率为输出向量,利用随机森林算法建立病穗率预测模型,为了避免高相关模型的偶然性,均重复建模 50 次,每次建模均随机抽取 3/4 的样本数作为训练样本、1/4 的样本数作为测试样本。由于不同区域的生育进程有所不同,苏南、苏中、苏北的起报时间也存在差异,起报时间不同使得预报时效也存在相应的差异,每年病穗率通常在当年 5 月末由植保站计算提供,根据最早的起报时间可以在 3 月初进行预测,意味着可以提前近 3 个月对病穗率进行预测,随着起报时间逐步向后推移,预测时效则逐步缩短,最短为提前 10 d。相关预测信息详见表 3、模型参数设置见表 4。

每一个起报时间,不同的 M_{try},通过重复建模 50 次,可以生成 50 个模型,不同模型对应的模拟精度存在差异,由于训练样本的模型精度均很高且相近,所以根据测试样本的模型精度来挑选最优随机森林模型,将筛选出的最优模型进行等权重集成,在一定程度上可以减少模型的随机误差和高相关的偶然性(徐敏等, 2017)。

不同起报时间最优模型模拟精度比较(表 5)发现:起报时间越接近乳熟期,随机森林模型模拟出的病

表 2 基于随机森林算法筛选出的影响病穗率的重要特征变量

Table 2 Important characteristic variables affecting diseased panicle rate of wheat scab based on RF algorithm

地区	越冬期至抽穗扬花期影响病穗率的重要特征变量	个数
苏南	抽穗扬花期(平均相对湿度、降水连续≥3 d的雨日总数、累计降水量、平均日日照时数、累计雨日、平均气温);拔节期(累计日照时数、平均相对湿度、累计降水量、最低气温≤0℃日数、降水量≥1 mm日数);越冬期(平均最高气温、雪深≥1 cm日数、降水量≥0.1 mm日数、累计最大积雪深度)	15
苏中	抽穗扬花期(平均相对湿度、降水连续≥3 d的雨日总数、平均日日照时数、累计降水量、平均气温);抽穗扬花前期(平均相对湿度、累计日照时数、降水量≥0.1 mm日数);拔节期(累计降水量、平均相对湿度、降水量≥1 mm日数、平均气温、降水量≥0.1 mm日数);越冬期(平均最高气温、雪深≥1 cm日数、累计降水量)	16
苏北	抽穗扬花期(平均相对湿度、平均日日照时数、降水连续≥3 d的雨日总数、平均气温、累计降水量、累计雨日);抽穗扬花前期(累计降水量、降水量≥0.1 mm日数、累计日照时数);拔节期(降水量≥0.1 mm日数、累计降水量、降水量≥1 mm日数、降水量≥5 mm日数、累计日照时数);越冬期(累计降水量、累计最大积雪深度、降水量≥0.1 mm日数)	17

表 3 利用随机森林算法建立病穗率预测模型的相关预测信息

Table 3 Information of diseased panicle rate prediction models established by RF algorithm

苏南			苏中			苏北		
起报时间	预报因子	预测时效	起报时间	预报因子	预测时效	起报时间	预报因子	预测时效
3月初	S1气象因子	提前近3个月	3月初	S1气象因子	提前近3个月	3月初	S1气象因子	提前近3个月
4月初	S1+S2气象因子	提前近2个月	4月初	S1+S2气象因子	提前近2个月	4月初	S1+S2气象因子	提前近2个月
4月下旬	S1+S2+S4气象因子	提前1个月	4月中旬	S1+S2+S3气象因子	提前40天	4月中旬	S1+S2+S3气象因子	提前40天
			5月上旬	S1+S2+S3+S4气象因子	提前20天	5月中旬	S1+S2+S3+S4气象因子	提前10天

注:S1:越冬期,S2:拔节期,S3:抽穗扬花前期,S4:抽穗扬花期。

表 4　随机森林建模过程中参数设置

Table 4　Specification of parameters for building the models using RF algorithm

苏南(病穗率样本数85)			苏中(病穗率样本数85)			苏北(病穗率样本数51)		
起报时间	M_{try}	N_{tree}	起报时间	M_{try}	N_{tree}	起报时间	M_{try}	N_{tree}
3月初	1, 2	600	3月初	1, 2	600	3月初	1, 2	600
4月初	2, 3, 4	600	4月初	2, 3, 4	600	4月初	2, 3, 4	600
4月下旬	3, 4, 5, 6, 7, 8	600	4月中旬	3, 4, 5	600	4月中旬	3, 4, 5	600
			5月上旬	4, 5, 6, 7, 8, 9		5月中旬	5, 6, 7, 8, 9, 10	
累计建模次数	550		累计建模次数	700		累计建模次数	700	

表 5　不同起报时间最优随机森林模型的预报准确率

Table 5　Forecast accuracy of optimal RF models initialized at different times

起报时间	训练样本模拟出的病穗率与实际病穗率的相关系数			测试样本预测出的病穗率与实际病穗率的相关系数		
	苏南	苏中	苏北	苏南	苏中	苏北
3月初	0.922**	0.916**	0.882**	0.748**	0.740**	0.698*
4月初	0.962**	0.965**	0.923**	0.817**	0.876**	0.759*
4月中旬		0.975**	0.924**		0.939**	0.763*
4月下旬	0.968**			0.855**		
5月上旬		0.977**			0.923**	
5月中旬			0.936**			0.765*
病穗率样本数	63	63	39	22	22	12

注: **通过了0.001的显著性t检验, *通过了0.01的显著性t检验。

穗率与实际病穗率的相关系数越高，说明在建立随机森林预测模型时，输入的影响病穗率的重要特征变量越多，则模型预测准确率越高；除了苏北地区训练样本的相关系数通过 0.01 显著性 t 检验以外，其余均通过了 0.001 显著性 t 检验，说明建立的随机森林模型具有较高的准确性；苏南和苏中地区，随机森林模型模拟出的病穗率与实际病穗率的相关系数高于苏北，这与赤霉病“南重北轻”的区域特征相关(徐敏等, 2019)，2002—2018 年，苏中、苏南、苏北年均病穗率分别为 23.0%、19.5%、8.5%，苏中和苏南病穗率超过 20.0% 的年份远远多于苏北，其中沿海地区是近年来赤霉病的重发生区域，在用随机森林算法进行数据挖掘时，赤霉病发生频次多的区域，更容易寻找病穗率与气象因子的非线性关系，即在每一棵决策树中更容易找寻出对病穗率影响大的变量，如果发生赤霉病的样本数很少，则较难捕捉病穗率与气象因子的对应关系。

3.3　不同生育期重要特征变量贡献率评价

为了解筛选出的各生育期重要特征变量在随机森林模型中的影响程度，计算各时段重要特征变量的贡献率。首先计算以苏南、苏中、苏北最迟起报时间建立的 6 个最优模型中各重要特征变量的 I_{MSE} 值占所有变量 I_{MSE} 累加值的比例；然后将 6 个模型中相同重要特征变量的比例进行平均；最后将属于同一生育期的变量权重进行累加，得到各生育期重要特征变量的贡献率。其中苏南地区由于抽穗扬花时间最早(从 4 月上旬开始)，因此不再单独计算抽穗扬花前期的贡献率。从图 2 可以看出，各生育期重要特征变量贡献率的排序为：抽穗扬花期(前期和高峰期)＞拔节期＞越冬期，说明抽穗扬花期的气象条件对最终的病穗率影响最大，起主导作用。苏

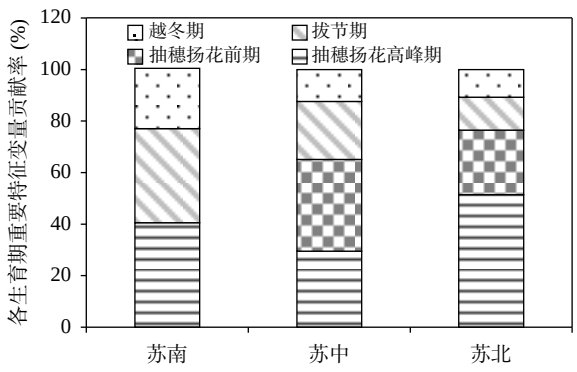

图 2　各生育期重要特征变量的贡献率

Fig. 2　Contribution rates of important characteristic variables during individual growth periods

南、苏中、苏北抽穗扬花期重要特征变量的贡献率分别为40.5%、65.1%、76.5%,其次是拔节期,越冬期对病穗率具有前期影响,影响程度相对弱一些;拔节期和越冬期重要特征变量的贡献率从苏南到苏北依次递减,这与抽穗扬花时间的早晚有关,苏南地区生育进程通常快于苏中和苏北,抽穗扬花时间与拔节期间隔最短,而苏北地区由于气温偏低,生育进程相对慢一些,抽穗扬花期要晚于苏中和苏南。

3.4 不同起报时间的最优随机森林模型的模拟验证

不同起报时间最优模型集成后的病穗率模拟值与实际病穗率进行对比(图3)发现:2002—2018年,13个地市不同起报时间的病穗率模拟值与实测值的波动趋势均一致,起报时间越接近乳熟期,模拟值总体越接近实测值,与表5得到的结论一致,说明随机森林模型对病穗率的预测具有较高的可靠性。病穗率模拟值波动幅度与实测值存在差异,低值区模拟值略偏大、高值区模拟值偏小,存在一定的系统性误差。因此,在具体使用过程中需要考虑这一特性。苏南、苏中、苏北最迟起报时间的病穗率模拟值与实测值的标准差分别是17.6%、21.3%、10.2%,标准差反映的是模拟值与实测值的偏差程度,说明随机森林模型对苏中的模拟误差大于苏南和苏北,主要是因为苏中病穗率≥40.0%的次数多于苏南和苏

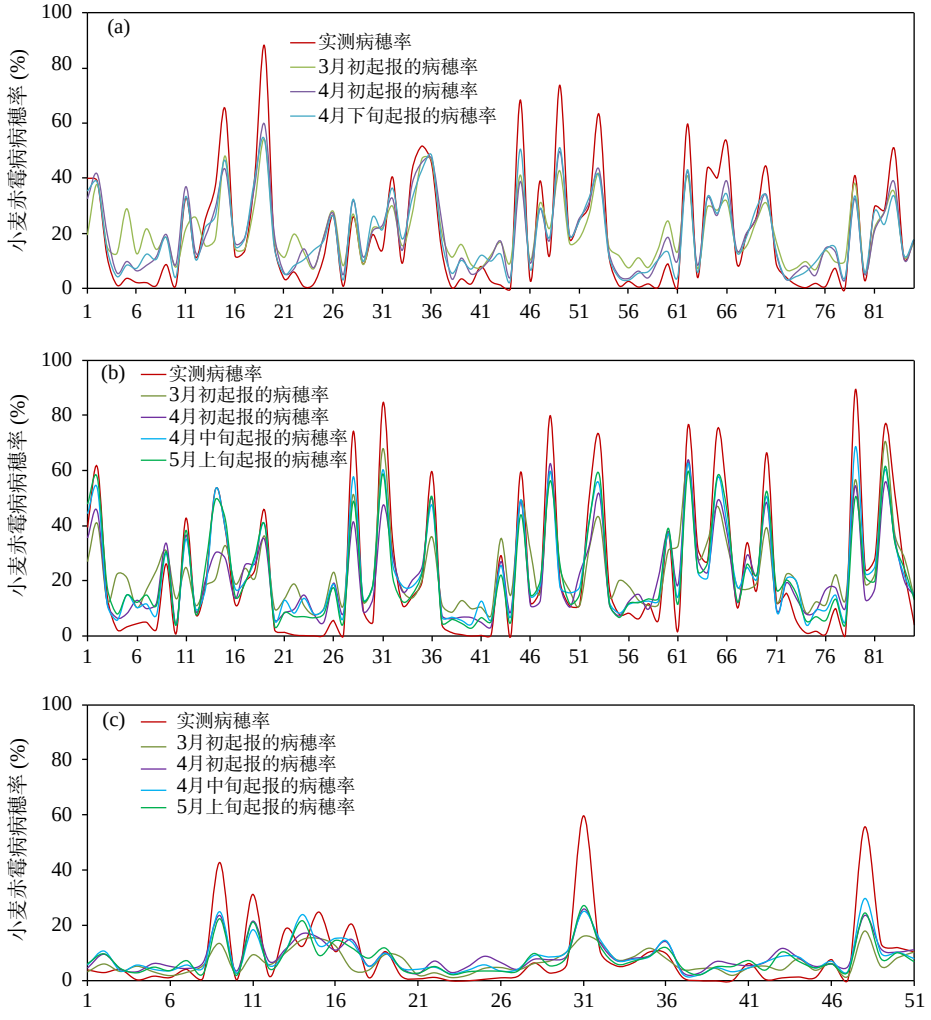

图3 2002—2018年不同起报时间随机森林最优模型集成后的病穗率模拟值与实测值的对比

(a. 苏南, b. 苏中, c. 苏北)

Fig. 3 Comparison of optimal model simulated rates of diseased panicle initialized at different times and the actual rates of diseased panicle during 2002 to 2018

(a. southern Jiangsu, b. central Jiangsu between the Yangtze and Huaihe Rivers, c. northern Jiangsu)

北，而模型对于高值的模拟偏小。

在农业气象业务服务中，通常通过赤霉病发生等级开展服务（王龙俊等，2017）。因此，需对模拟出的病穗率进行等级划分，做进一步的验证。按照国家标准 GB/T 15796-2011《小麦赤霉病测报技术规范》规定的赤霉病发生程度分级指标，赤霉病的发生程度可分为 6 级，即 0 级（未发生）、1 级（轻发生，病穗率 0.1%—10.0%）、2 级（中等偏轻发生，病穗率 10.1%—20.0%）、3 级（中等发生，病穗率 20.1%—30.0%）、4 级（偏重发生，病穗率 30.1%—40.0%）、5 级（大发生，病穗率≥40.1%）。从图 4 可以看出，最迟起报时间建立的随机森林最优模型集成后的模拟等级与实际等级的空间分布总体一致，赤霉病发生程度均为“南强北弱”，尤其是沿海和苏南地区，说明随机森林模型的分级结果能揭示出赤霉病总的空间格局和内在规律，淮北地区由于本身赤霉病发生程度较轻，2002—2018 年宿迁、徐州、连云港等 3 市均仅有 1 次达到 5 级，“大发生”样本数太少，因此未能模拟出。统计结果表明，苏南、苏中、苏北最优模型集成后的赤霉病等级与实际赤霉病等级完全一致的准确率分别是 62.4%、64.7%、62.7%，偏差一级的分别占 34.1%、32.9%、25.5%，偏差两级的分别占 3.5%、2.4%（含三级）、11.8%，其中偏差一级的主要集中在“轻发生”和“大发生”，与徐敏等（2019）利用赤霉病综合影响指数判定江苏全省赤霉病等级完全一致的准确率为 43.0% 相比，随机森林模型的准确率明显提高，说明随机森林模型对赤霉病等级的模拟能力总体较好。在实际应用中，当预测等级为“大发生”时，需要格外注意，因为实际将发生的等级很可能比预测的要严重。由此表明随机森林最优模型在赤霉病等级模拟方面同样具有较好的适用性，为建立赤霉病发生等级预测模型提供了新思路。

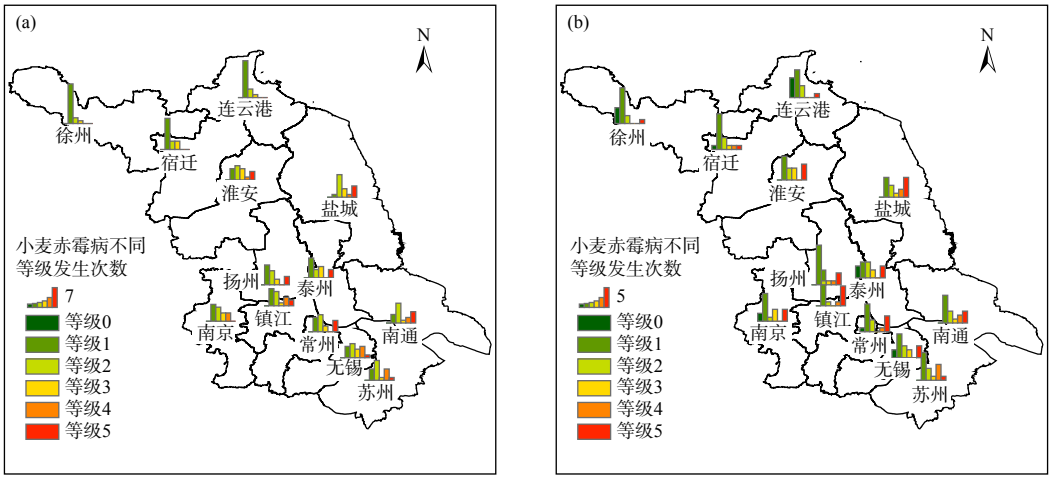

图 4　2002—2018 年最迟起报时间基于随机森林最优模型模拟出的小麦赤霉病各等级次数与实际等级次数

（a. 模拟等级，b. 实际等级）

Fig. 4　Simulated number of wheat scab levels based on RF model with the latest initial time and actual number of wheat scab levels from 2002 to 2018

（a. simulated number, b. actual number）

4　结论与讨论

以江苏小麦赤霉病病穗率为研究对象，利用随机森林机器学习算法，以精度平均减少值为评价指标，结合赤霉病菌的病理，分生育期、分区域筛选出对病穗率影响相对重要的特征变量，然后根据不同的起报时间，通过训练样本和测试样本的多次学习，选取最优预测模型，并进行模型模拟精度的验证。得到以下主要结论：

（1）随机森林算法重要性度量表明，小麦在不同的生育阶段，对赤霉病菌产生影响的气象因子不同，越冬期主要是气温和降雪；拔节期主要是日照、降雨量、湿度和雨日；抽穗扬花期主要是湿度、降水连续≥3 d 的雨日和日照。甄别出的重要特征变量排序结果符合赤霉病菌发育、释放、侵染和流行的生理学规律。

(2)苏南、苏中、苏北抽穗扬花期重要特征变量的贡献率分别为40.5%、65.1%、76.5%,该生育期的气象条件对最终的病穗率影响最大,具有决定性作用;拔节期气象条件的影响程度位列第二;越冬期气象条件的影响程度相对较弱。拔节期和越冬期的气象条件对病穗率具有前期影响。

(3)苏南、苏中、苏北开始进行病穗率预测的时间最早可在3月初,最迟预测时间分别是4月下旬、5月上旬、5月中旬,预测时效最长可达近3个月,起报时间越接近乳熟期,输入的重要特征变量越多,病穗率预测准确率也随之越高。经过检验,模型对病穗率时间波动特征模拟的非常好,对赤霉病"中等"和"偏重"等级模拟的也不错。

不同起报时间的最优随机森林模型对于"大发生"的模拟均过于"保守",模拟值均低于实际值,可能与考虑的特征变量还不够全面有关,因为赤霉病的发生、发展不仅与气象条件密切相关,还与秸秆持续还田、小麦种植品种、氮肥使用量、田间管理措施等因素有关(李韬等,2016)。随着小麦生育期的推进,可在不同的关键时间节点开展病穗率预测,文中的建模思路为动态预测病穗率提供了新的方法和思路,但由于预报因子采用的是时段平均(月或旬时间尺度),还未达到真正意义上的动态预测的时间精度,在今后的研究中可以考虑细化预报因子时间尺度,重新利用随机森林算法建模,以期进一步提高预测准确率。

综合而言,随机森林机器学习算法可在病穗率预测中进行应用,建立的预测模型具有较高的可靠性和准确性。具有较长预测时效的预测结果可为植保部门分区治理、统防统治、适时预防提供指导,为广大农户提前购买农药的数量和开展化学防治提供充裕的准备时间,为最大限度减轻病害流行和危害程度提供可能,助力农药减施增效,为保护农田生态环境奠定基础。

参考文献

刁春友, 朱叶芹. 2006. 农作物主要病虫害预测预报与防治. 南京: 江苏科学技术出版社, 389pp. Diao C Y, Zhu Y Q. 2006. Forecast and Control of Main Crop Diseases and Insect Pests. Nanjing: Jiangsu Scientific and Technical Press, 389pp (in Chinese)

侯明生, 黄俊斌. 2006. 农业植物病理学. 北京: 科学出版社, 480pp. Hou M S, Huang J B. 2006. Agricultural Plant Pathology. Beijing: Scientific Press, 480pp (in Chinese)

黄冲, 姜玉英, 吴佳文等. 2019. 2018年我国小麦赤霉病重发特点及原因分析. 植物保护, 45(2): 160-163. Huang C, Jiang Y Y, Wu J W, et al 2019. Occurrence characteristics and reason analysis of wheat head blight in 2018 in China. Plant Protection, 45(2): 160-163 (in Chinese)

霍治国, 王石立. 2009. 农业和生物气象灾害. 北京: 气象出版社, 285pp. Huo Z G, Wang S L. 2009. Agricultural and Biometeorological Disasters. Beijing: China Meteorological Press, 285pp (in Chinese)

贾金明. 2002. 黄河中下游小麦赤霉病气象指数的建立与应用. 气象, 28(3): 50-53. Jia J M. 2002. Development and application of wheat scab weather index in Yellow River Basin. Meteor Mon, 28(3): 50-53 (in Chinese)

姜明波, 翟顺国, 王守强等. 2018. 信阳地区小麦赤霉病发生与春季降水的相关性分析. 河南农业科学, 47(11): 80-84. Jiang M B, Zhai S G, Wang S Q, et al. 2018. Correlation between occurrence of fusarium head blight of wheat and spring rainfall in Xinyang area. J Henan Agric Sci, 47(11): 80-84 (in Chinese)

赖成光, 陈晓宏, 赵仕威等. 2015. 基于随机森林的洪灾风险评价模型及其应用. 水利学报, 46(1): 58-66. Lai C G, Chen X H, Zhao S W, et al. 2015. A flood risk assessment model based on random forest and its application. J Hyd Eng, 46(1): 58-66 (in Chinese)

李韬, 郑飞, 秦胜男等. 2016. 小麦-黑麦易位系T1BL•1RS在小麦品种中的分布及其与小麦赤霉病抗性的关联. 作物学报, 42(3): 320-329. Li T, Zheng F, Qin S N, et al. 2016. Distribution of wheat-rye translocation line T1BL •1RS in wheat and its association with fusarium head blight resistance. Acta Agron Sinica, 42(3): 320-329 (in Chinese)

陆维忠. 2001. 小麦赤霉病研究. 北京: 科学出版社, 156pp. Lu W Z. 2001. Study on Wheat Scab. Beijing: Scientific Press, 156pp (in Chinese)

石礼娟, 卢军. 2017. 基于随机森林的玉米发育程度自动测量方法. 农业机械学报, 48(1): 169-174. Shi L J, Lu J. 2017. Automatic measurement method for maize ear development degree based on random forest. Trans Chin Soc Agric Mach, 48(1): 169-174 (in Chinese)

王超, 阚瑷珂, 曾业隆等. 2019. 基于随机森林模型的西藏人口分布格局及影响因素. 地理学报, 74(4): 664-680. Wang C, Kan A K, Zeng Y L, et al. 2019. Population distribution pattern and influencing factors in Tibet based on random forest model. Acta Geogra Sinica, 74(4): 664-680 (in Chinese)

王利民, 刘佳, 杨玲波等. 2018. 随机森林方法在玉米-大豆精细识别中的应用. 作物学报, 44(4): 569-580. Wang L M, Liu J, Yang L B, et al. 2018. Application of random forest method in maize-soybean accurate identification. Acta Agron Sinica, 44(4): 569-580 (in Chinese)

王龙俊, 丁艳峰, 郭文善等. 2017. 农事实用旬历手册. 3版. 南京: 江苏凤凰科学技术出版社, 100pp. Wang L J, Ding Y F, Guo W S, et al. 2017. Handbook of Every Ten Days Calendar for Agricultural Activities. 3rd ed. Nanjing: Jiangsu Phoenix Scientific and Technical Press, 100pp (in Chinese)

王晓曦, 王修法, 温纪平等. 2008. 世界小麦产量及加工业发展概况. 粮食加

工, 33(4): 11-12, 18. Wang X X, Wang X F, Wen J P, et al. 2008. Overview of world wheat yield and processing industry development. Grain Process, 33(4): 11-12, 18 (in Chinese)

吴春艳, 李军, 姚克敏. 2003. 小麦赤霉病发病程度的预测. 中国农业气象, 24(4): 19-22. Wu C Y, Li J, Yao K M. 2003. Prediction of damage level of scab of wheat in Shanghai. Chinese J Agrometeorol, 24(4): 19-22 (in Chinese)

吴孝情, 赖成光, 陈晓宏等. 2017. 基于随机森林权重的滑坡危险性评价: 以东江流域为例. 自然灾害学报, 26(5): 119-129. Wu X Q, Lai C G, Chen X H, et al. 2017. A landslide hazard assessment based on random forest weight: A case study in the Dongjiang River Basin. J Nat Disast, 26(5): 119-129 (in Chinese)

肖晶晶, 霍治国, 李娜等. 2011. 小麦赤霉病气象环境成因研究进展. 自然灾害学报, 20(2): 146-152. Xiao J J, Huo Z G, Li N, et al. 2011. Progress in research on meteorological conditions of wheat scab. J Nat Disast, 20(2): 146-152 (in Chinese)

徐敏, 高苹, 刘文菁等. 2017. 水稻稻曲病气象等级预报模型及集成方法. 江苏农业科学, 45(17): 95-98. Xu M, Gao P, Liu W J, et al. 2017. Study on meteorological grade prediction models and integration methods of rice false smut. Jiangsu Agric Sci, 45(17): 95-98 (in Chinese)

徐敏, 高苹, 徐经纬等. 2019. 江苏省小麦赤霉病综合影响指数构建及时空变化特征. 生态学杂志, 38(6): 1774-1782. Xu M, Gao P, Xu J W, et al. 2019. Construction of the comprehensive impact index for wheat scab and its spatiotemporal variations in Jiangsu Province. Chinese J Ecol, 38(6): 1774-1782 (in Chinese)

曾娟, 姜玉英. 2013. 2012 年我国小麦赤霉病暴发原因分析及持续监控与治理对策. 中国植保导刊, 33(4): 38-41. Zeng J, Jiang Y Y. 2013. Analysis on the causes of wheat scab outbreak in China in 2012 and its sustainable monitoring and control countermeasures. China Plant Prot, 33(4): 38-41 (in Chinese)

张汉琳. 1987. 气象因素与麦类赤霉病群体流行波动的研究. 气象学报, 45(3): 338-345. Zhang H L. 1987. A study of the fluctuation of the cereal's gibberellin prevalence and the weather factors in the middle and lower Chang-Jiang. Acta Meteor Sinica, 45(3): 338-345 (in Chinese)

张雷, 刘世荣, 孙鹏森等. 2011. 气候变化对马尾松潜在分布影响预估的多模型比较. 植物生态学报, 35(11): 1091-1105. Zhang L, Liu S R, Sun P S, et al. 2011. Comparative evaluation of multiple models of the effects of climate change on the potential distribution of *Pinus massoniana*. Chinese J Plant Ecol, 35(11): 1091-1105 (in Chinese)

张雷, 刘世荣, 孙鹏森等. 2011. 气候变化对物种分布影响模拟中的不确定性组分分割与制图: 以油松为例. 生态学报, 31(19): 5749-5761. Zhang L, Liu S R, Sun P S, et al. 2011. Partitioning and mapping the sources of variations in the ensemble forecasting of species distribution under climate change: A case study of Pinus tabulaeformis. Acta Ecol Sinica, 31(19): 5749-5761 (in Chinese)

Han J W, Kamber M 著. 范明, 孟小峰译 .2007. 数据挖掘: 概念与技术. 北京: 机械工业出版社, 488pp. Han J W, Kamber M. Fang M, Meng X F, trans. 2007. Data Mining: Concepts and Techniques. Beijing: Machinery Industry Press, 488pp (in Chinese)

Biau G. 2012. Analysis of a random forests model. J Machine Learning Res, 13(1): 1063-1095

Breiman L. 2001. Random forests. Mach Learn, 45(1): 5-32

Champeil A, Doré T, Fourbet J F. 2004. Fusarium head blight: Epidemiological origin of the effects of cultural practices on head blight attacks and the production of mycotoxins by Fusarium in wheat grains. Plant Sci, 166(6): 1389-1415

Donnelly S, Walsh D. 1996. Quality of life assessment in advanced cancer. Palliat Med, 10(4): 275-283

Iverson L R, Prasad A M, Matthews S N, et al. 2008. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. Forest Ecol Manage, 254(3): 390-406

Starkey D E, Ward T J, Aoki T, et al. 2007. Global molecular surveillance reveals novel *Fusarium* head blight species and trichothecene toxin diversity. Fungal Genet Biol, 44(11): 1191-1204

Verikas A, Gelzinis A, Bacauskiene M. 2011. Mining data with random forests: A survey and results of new tests. Patt Recognit, 44(2): 330-349