**分类号：S431. 2 授予学位单位代码: 10434**

**研究生学号 : 2015120181**

**山東農業大學**

**全日制硕士专业学位论文**

基于R语言的麦田灰飞虱监测预警模型构建

**Establishment of Monitoring and Forecasting**

**Model of *Laodelphax striatellus* in the Wheat Field**

**Based on R Program**

|  |  |  |
| --- | --- | --- |
| **姓名** | ： | **汪深** |
| **学位类别** | **：** | **农业推广** |
| **专业** | **：** | **植物保护** |
| **研究方向** | **：** | **农业大数据** |
| **学院** | **：** | **植物保护学院** |
| **指导教师** | **：** | **刘勇 教授** |

**2017**年**6**月**7**日

论 文 提 交 日 期 ： 2016年4月

论 文 答 辩 日 期 ： 2016年6月

学 位 授 予 日 期 ： 2016年6月

学 科 门 类 ： 农 学

答辩委员会主席 ： 张礼生教授

关于学位论文原创性和使用授权的声明

本人所呈交的学位论文，是在导师指导下，独立进行科学研究所取得的成果。对在论文研究期间给予指导、帮助和做出重要贡献的个人或集体，均在文中明确说明。本声明的法律责任由本人承担。

本人完全了解山东农业大学有关保留和使用学位论文的规定，同意学校保留和按要求向国家有关部门或机构送交论文纸质本和电子版，允许论文被查阅和借阅。本人授权山东农业大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文，同时授权中国科学技术信息研究所将本学位论文收录到《中国学位论文全文数据库》，并向社会公众提供信息服务。

保密论文在解密后应遵守此规定。

论文作者签名：

导 师 签 名 ：

日 期 ：

目 录

**中 文 摘 要 ...I**

**Abstract II**

**1 前言 1**

1.1 大数据时代的到来 1

1.1.1 大数据的发展 1

1.1.2 我国大数据的发展现状 2

1.2 大数据在植物保护中的应用 3

1.2.1 农业大数据 3

1.2.2 大数据在植物保护中的应用 4

1.2.3 农业大数据的面临的挑战 4

1.3 数据挖掘的方法 5

1.3.1 决策树算法 6

1.3.2 随机森林算法 7

1.4 灰飞虱的监测预警 8

1.4.1灰飞虱监测预警的重要性 9

1.4.2 温度对灰飞虱发生程度的影响 10

1.4.3湿度对灰飞虱发生程度的影响 11

1.4.4 风速对灰飞虱的影响 11

1.4.5国内灰飞虱已建立的监测预报模型 12

1.5 立题依据与意义 13

**2 材料与方法 15**

2.1 数据来源 15

2.1.1 气象数据来源 15

2.1.2 灰飞虱数据来源 15

2.2 灰飞虱发生程度分级及因变量分级 15

2.3 自变量的选择 16

2.4 随机森林算法运行原理 16

2.4.1 决策树算法运行原理 17

2.4.2 Bagging集成学习算法 17

2.4.3 袋外估计理论 18

**3 结果 19**

3.1 数据的整理导入及调用 19

3.2 mtry值的确定 20

3.3 ntree值的确定 21

3.4构建监测预警模型并分析主要影响因子 22

3.5 模型准确率的检验 24

3.6 使用监测预警模型进行预测 27

**4 讨论 28**

4.1 风速与降水对灰飞虱发生程度的影响 28

4.2 模型对于其他地区或其他昆虫的应用分析 28

4.3 研究的优势及展望 29

**5 结论 31**

**创新点 32**

**参考文献 33**

**致 谢 38**

**附 录 39**

中 文 摘 要

近年来，大数据逐渐走入我们的生活，在各行各业的应用中都有所涉及。而农业作为基础产业与大数据结合，使得现代农业的发展步入新领域。植物保护与大数据的结合，是农业大数据中重要的一部分。利用长周期的大田调查数据与气象数据相结合，选择相应的数据分析算法并利用特定的数据分析平台进行运算，分析结果并进行预测，是农业大数据的一种基本流程。作为我国常见的粮食害虫，灰飞虱（*Laodelphax striatellus*）广泛分布于包括山东省在内的我国各个地区，自2008年爆发以来，迅速成为山东地区主要危害虫类，使山东粮食的安全生产蒙受了重大损失。所以，急需建立针对山东地区的灰飞虱监测预警模型，探究对于其暴发成灾起主要作用的影响因子，并对今后灰飞虱的发生状况进行预测，从而能够实现灰飞虱的爆发提前预警，预先做好防治措施，降低其对粮食作物的危害程度。使用随机森林算法对2008年 ~ 2016年共计71次的田间调查数据进行整理，以R语言为实现平台，以“累计30天平均、累计60天平均、累计90天平均”三种时间尺度合计33种气象因子为自变量，以济宁地区的灰飞虱发生程度为因变量，构建监测预警模型，找出影响灰飞虱发生的主要气象因子，从而对灰飞虱的发生程度进行预测预报。通过模型分析显示，对济宁地区灰飞虱发生程度起主要作用的气象因子依次有“累计60天平均最大风速”和“累计90天平均最大风速”，有一定影响作用的气象因子依次有“累计60天平均降水量”和 “累计30天平均降水量”。利用随机森林模型袋外估计进行检验，模型正确率达到88.89%，利用测试数据进行检验，模型正确率达到88.24%，随机森林监测预警模型的预测准确率较高，具有科学性和实用性。可用于实际监测预警，为济宁乃至山东地区的植保工作开展提供参考。

关键词：农业大数据；灰飞虱；监测预警；R语言；随机森林；气象因子

## **Abstract**

In recent years, big data gradually into our lives, in all sectors of the application are involved. And agriculture as a basic industry and big data combination, making the development of modern agriculture into new areas. The combination of plant protection and big data is an important part of agricultural big data. Using the long-term field survey data and meteorological data, selecting the corresponding data analysis algorithm and using the specific data analysis platform to carry out the operation, analyze the results and forecast, is a basic process of agricultural data.As a common food pest in China, *Laodelphax striatellus* is widely distributed in various regions of China, including Shandong Province. Since the outbreak in 2008, it has become the main hazard insect in Shandong province. The food security poses a great threat. Therefore, it is urgent to establish a model for the monitoring and forecasting of *Laodelphax striatellus* in Shandong area, to explore the influencing factors of the main factors of its outbreak and to predict the occurrence of *Laodelphax striatellus* in the future, so as to realize the early warning of the outbreak of *Laodelphax striatellus* , In advance to prevent and control measures to reduce the extent of its harm to food crops.A total of 71 field data from 2008 to 2016 were compiled using R-based and Rstudio-based random forest algorithm. The data were analyzed for the data of the occurrence of *Laodelphax striatellus*, and the cumulative 30 days The average meteorological factors were measured, and the main meteorological factors affecting the occurrence of *Laodelphax striatellus* were determined by using the three kinds of meteorological factors as the independent variables, and the degree of occurrence of *Laodelphax striatellus* in Jining area as the dependent variable, So as to predict the occurrence degree of *Laodelphax striatellus*.The results showed that the meteorological factors, which had the main effect on the occurrence of *Laodelphax striatellus* in Jining area, were "cumulative 60 days average maximum wind speed" and "cumulative maximum daily wind speed of 90 days", and the meteorological factors with certain influence were 60 days average precipitation "and" accumulated 30 days average precipitation ". The correctness rate of the model is 88.89%, and the correctness of the model is 88.24% by using the test data, the prediction accuracy of the random forest monitoring and forecasting model is high and scientific and practicable. Can be used for practical monitoring and early warning, for the Jining and even Shandong area of the plant protection work to provide a reference.

**Key words: Agricultural big data; *Laodelphax striatellus*; Monitoring and Forecasting; R Program; Random Forest; Meteorological factors**

1 前言

1.1 大数据时代的到来

1.1.1 大数据的发展

近年来，大数据在人们视野中出现的频次越来越高，在我们尚未意识到之际，大数据时代已悄然到来。在我党的十八大之后，更是将发展大数据写入我国“创新驱动发展战略”之中，从而使大数据越来越为我们所认识、接受，可以说，在我国发展“大数据+相关产业”的最好时期已经来临。而随着“大数据+相关产业”的高速发展，大数据的应用也开始逐渐渗透入生活的各个方面，为了我们的生活和工作带来了很多的便利。可以说，数据，已经渗透到当今每一个行业和业务职能领域，成为一种重要的生产因素，人们对于大量数据的挖掘和应用，使得我们正逐步迈入大数据时代。

大数据也被成为巨量资料或巨量数据，它是指在一定长度的时间内无法通过常规软件工具对目标进行抓取、挖掘、管理和处理的数据集合。IBM最早定义了大数据的四个特征，我们称之为“4V”，即：规模性（Volume） 、多样性（variety）、价值性（Value）以及高速性（velocity）。这四个“V”属性也体现了大数据在四个方面的特征：第一是数据体量巨大。数据量已从TB级别跃升到PB、EB级别，乃至于当今的ZB级别；第二是数据类型繁多，数据的表现形式多种多样，并不拘泥于一种表现形式，如网络文章、视频资料、图片信息、访问记录、地里位置信息等；第三是价值密度低但价值量高，只要能够使用合适的处理方法对数据进行正确、准确的分析，就能挖掘很高的汇报价值；第四是处理速度快，可以利用合理的相关算法，基于相关平台进行各种类型的数据的快速处理从而快速获得高价值的信息，这一点和传统的数据挖掘技术有着本质的区别。

“大数据”这个概念的首次提出实在1980年，未来学大师阿尔文•托夫勒在其著作《第三次浪潮》中将大数据称为“第三次浪潮的华彩乐章”（温孚江，2013）。之后的几年里，来自美国、匈牙利等多个国家的科学家，越来越意识到收集数据的重要性以及数据的背后蕴藏的战略意义，同时针对大数据相关方面的探索与研究活动也越来越多。2008年，Google公司成立10周年，在此之际著名的《Nature》杂志出版了一期专刊，在专刊中对于大数据处理未来的应用前景以及相关的技术难题和挑战进行了讨论，专刊中使用的正是“Big - Data”一词（陈颖，2015）。直到2009年，“大数据”一次逐渐成为互联网以及信息技术行业中的热门词汇，“大数据“一词又再次活跃于人们眼中，这次它的发展呈星火燎原之势迅速升温并席卷全球，美国更是在2014发布了关于关于全球大数据的白皮书研究报告——《大数据：抓住机遇、守护价值》，提出使用大数据技术构建相关的政策框架，将大数据应用于公民医疗保健、公民隐私、国土安全、企业发展、消费者服务、社会反歧视方面。

1.1.2 我国大数据的发展现状

我国在大数据方面的研究虽然起步较晚，但其发展呈后来居上之势，进展迅速，国家政府部门以及相关企业也对此给予了高度关注。

2011年12月，国务院工信部就物联网建设在“十二五”规划之中提出了四项关键创新工程，其中就包括信息处理技术，内含海量数据存储、数据挖掘、图像视频智能分析，这些都是大数据中的关键核心技术，从而将建设大数据上升到国家战略的高度。2016年3月发布的“十三五”规划纲要中指出，实施国家大数据战略。把大数据作为基础性战略资源，全面实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助力产业转型升级和社会治理创新。

为了更好的推进国家战略的实施，我国启动了国家级超大云数据中心项目，分别在北京市、贵州贵阳市、内蒙古乌兰察布市成立了以首都为中心，由南到北的三大国家大数据中心。三大中心的成立 ，实现了国家旅游局北京机房与贵州灾备中心数据的同步传输和异地备份，并为华北地区、京津地区的提供了国家级云计算服务，为承接高科技产业、加快产业转型升级提供强有力的支撑(黄锋，2016)。

我国互联网企业中的代表——阿里巴巴集团，为了挖掘数据蕴含的价值，专门在管理层设立了“首席数据官”这一职位，并成立了大型数据分享平台——“聚石塔”，使得“数据分享战略”得以全方位推进，为其旗下各个购物网站中的电商、电商服务商提供了数据云服务，这也使得阿里巴巴成为最早提出通过数据分享使业务进行数据化运营的企业（唐洪明，2014）。

据相关的统计资料显示，截至2016年12月，我国网民规模达7.31亿，普及率达53.2 % ，超过全球平均水平3.1个百分点，超过亚洲平均水平7.6个百分点。全年共计新增网民4299万人，增长率为6.2 % 。中国网民规模已相当于欧洲人口总量（第39次中国互联网络发展状况统计报告，2017）。网络用户数量的大量增加，为“大数据 + 网络相关产业”的发展提供了发展的基础和保障，他们每天的活动将产生庞大的信息、数据，用户的习惯、爱好、活动轨迹、搜索内容等各种数据都包含极大的价值。大数据的发展与互联网的发展相辅相成，互联网一刻不停地在为大数据提供海量的数据支持，而大数据引领互联网以及更多的其他领域的发展方向。所以目前在我国互联网相关产业飞速发展之际，把传统领域与大数据分析相结合，将成为一种有极大前景的发展方向。

1.2 大数据在植物保护中的应用

1.2.1 农业大数据

目前大数据与其他领域的合作正在如火如荼地开展着。例如:在金融行业，已经实现了在信用支付，通过调取针对某个用户的相关资料与征信记录，综合分析该用户的信用历史、行为方面的偏好、履约能力的大小、身份特质、人脉关系等信息，直接与其信用额度关联，构筑信用消费生态；在零售行业，通过分析消费者的多次购买记录，实现了对消费者主要需求、购买能力大小、购买频次、不同产品间的购买搭配等多种数据的挖掘，再结合针对消费者个人信息的分析，来探究不同年龄、不同层次、不同消费水平的消费者的爱护与需求，从而进行有针对性的货物摆放及广告推送；在政治领域，政府各部门通过网上政务平台以及微信公众号、官方微博等方式与民众进行互动，并从中收集民众关注的民生热点、议论热点，来为下一步决策提供指导以及了解已推行政策的执行和落实程度，从而提高政务处理的效率与能力；在气象领域，可以使用观测积累的气象数据构建相关的短期预测预报模型，进行不同地区的高准确度短期天气预报预测，或是构建适用于更长时间尺度的分析预测模型，对不同面积大小的地区进行气候预测，并针对可能出现的大范围气象灾害提前进行准备。

可以说大数据虽然大数据的应用与发展，在我国起步不久，但在各个行业领域内都得到了有效的利用并展示出很好的表现，极大地方便、改善了我们的生活。而在农业领域，大数据与传统农业结合的相关研究程度还涉足不深，研究覆盖方面也不够完全，而在却有着较大的研究前景与实用价值。一方面，在传统农业中现代化的科技力量不强，技术水平不高，亟待提升。比如在作物品种一定的之前提下为了保证产量只好进行大量的农药喷洒、化肥施用、水灌溉，不仅造成了资源的大量浪费以及环境污染，也对我们的食品安全产生了威胁，而将相关的数据进行积累、收集，可以通过数据挖掘找到合适的农药、化肥的施用量与水的灌溉量，从而在保证粮食产量的前提下减少对环境及食品安全的破坏。另一方面，大量的数据可供我们进行分析挖掘。农业产品的质量、产量受到诸如：气候、地理环境、土壤质量、人类行为等多种因素的共同影响，农产品的生产、加工、储藏、运输、销售等各个环节涉及多个行业，这其中可以产生大量有价值的数据，通过对这些数据的分析挖掘可以从中提取出大量商业价值、生态价值、社会价值。

当前我国农业领域还存在一些问题，如食品安全、土壤污染治理、病虫害预测与防治、动植物育种、产业结构调整、农产品价格、农副产品消费、小城镇及乡村建设等领域，都可以通过大数据的应用研究进行预测和指导。将大数据研究与农业领域的相关科学研究进行结合，不仅可以为农业科研、提供新方法，还能够为政府决策、涉农企业发展等提供新的思路（温孚江，2013）。

1.2.2 大数据在植物保护中的应用

大数据在植物保护中应用属于大数据在农业中应用的一个分支。植物保护是农业中生产不可缺少的组成部分。植物保护相关工作中的核心就是对农作物的主要病虫害进行监测预警，它对指导农作物进行病虫害的科学防控、保障我们的粮食和农产品安全有着重要的作用。

目前针对病虫害监测预警方面的研究方法主要是使用已有一些专用的统计分析技术和软件，虽然在一定程度上具有科学性，但是在研究过程中往往存在采集数据指标偏少的情况，造成预测结果不够准确，与病虫害的实际发生量匹配度不高。因此依靠对多年农作物病虫害发生期、作物生长发育状况、气候条件、周边自然环境、土壤特点、农业管理措施等数据的收集、整合，以大数据的历年、处理技术及分析方法进行数据挖掘、构建监测预警模型，可以针对病虫害的发生趋势及程度进行监测预警，这将是植保研究的一种未来发展趋势（杨波等，2014）。

1.2.3 农业大数据的面临的挑战

虽然在目前，不管是国际方面还是国内方面，不管是中央政府还是地方政府，都对大数据的发展尤为重视，农业大数据作为大数据中重要的一个分支也遇到了前所未有的发展机遇，但是也不能否认在农业大数据的实际研究过程中还存在一些挑战。

首先，农业大数据的发展不是依靠一个学科的人才就可以完成的，在农业大数据团队的构建过程中不仅需要农业方面的人才，还需要数据分析方面的人才，来共同完成这项跨专业、跨领域的研究。而目前对于农业大数据的研究尚处于起步阶段，所以这方面配合研究的经验比较少，缺乏可参考的模式，需要不断地进行摸索。

其次，农业各方面活动中虽然能够产生大量的数据，但是之前对于相关数据的系统化、程序化、模式化的收集行为较少，大量富有价值的数据在无意中流失殆尽，而在农业相关数据的收集活动中，对于收集内容的选择、收集方法的选择、收集人员的调配、收集周期的把控也都没有形成一个合理、高效、科学的流程和模式，这在接下来今后的研究工作中都需要进行注意，也需要摸索。

再者，农业数据的整理和分析过程较为繁重，需要将各种类型的数据，如文本、图像、高维数据等类型数据，先降低纬度后再进行度量和处理，直至输出为便于统计的数据类型，并进行格式的规范；在进行农业数据的分析时，由于结合气象、土壤、人类活动、历年产量等各方面的数据，所以在对于方法进行选择时，要考虑不同分析算法之间的差别、优劣、擅长数据类型等多个方面，从而选择最具有科学性、实用性的分析方法，保证分析结果的准确性。

最后，由于农业大数据的调查、分析涉及多个层面，所以不仅仅是一项科学研究、农业工作，更是一项社会活动。但是目前在政策法律、数据保密、数据共享、人才培养及调配等方面都存在着严峻的挑战，这需要政府农业部门、相关企事业单位及个人的共同努力，来为大数据在农业中的应用提供社会层面的支持（谢润梅，2015）。

1.3 数据挖掘的方法

数据挖掘就是运用科学的数据分析方法，从海量的数据中提取我们需要的信息的过程，是一个将大数据去粗取精、去伪存真的过程。单纯的海量数据仅仅是蕴含价值，而只有通过数据挖掘，才能实现海量数据中的价值。所以，数据的分析挖掘是大数据处理中的关键环节。

为了实现数据挖掘的环节，则需要使用机器学习。机器学习是农业大数据中分析和挖掘的有效手段，是通过训练及优化构建最符合样本统计规律的预测分析模型，来帮助人们揭示实物发生和发展的基本规律，从而发现数据与数据间隐藏相关性的过程。通过机器学习，可以探索出数据之中隐藏的多种相关性，从而准确高效地预测判断事物的发生规律和发展方向，指导人们做出更为符合实际规律的决策（Baskshi *et a*l*.*, 1993）。

目前常用的大数据挖掘技术主要有决策树算法、BP人工神经网络、遗传算法、支持向量机、随机森林算法等（Breiman *et al*.，1996；孙增圻，1996；王光宏等，2004；邹祎，2016）。本文实验中使用的是决策树算法及随机森林算法进行数据分析。

1.3.1 决策树算法

决策树（decision tree）是一个种类似于流程图的树结构的发展较为成熟的算法，由多个分枝结构组成，其中每个内部节点表示一个属性上的测试，每个分枝代表一个测试输出，而每个叶节点代表一种类别（Quinlan, 1986）。决策树于上世纪八十年代由Breiman等人提出（Breiman et al.，1984），是通过反复二分数据来进行数据的分类或回归，它的目标也是建立一种回归或分类模型，由于它输出结果的展示方式像一棵倒置的树，所以被称为决策树。

决策树算法这一数据挖掘方法起源于概念学习系统（Concept Learning System CLS）。之后，在CLS的基础上发展到经典的ID3算法，后演化为C4.5算法，其运算函数与ID3算法有了很大改进，效果更好，目前的C5.0较C4.5在运行速度上有了极大的提高，适合用于大数据集的处理，后来在C4.5的基础上又演化为能处理连续变量的更高级算法，也就是C5.0算法。经过多次改进后决策树算法已经相对成熟（Quinlan, 1993）。

在数据挖掘中使用决策树算法的优点是决策的制定过程是可见的，不需要花费较长时间进行构造过程，其描述简单直观，易于用户理解，同时分类速度快，可以节省运算的时间，所以决策树法擅长处理非数值型数据，例如分类数据是它的擅长领域，且由于其分类速度较快，特别适合大规模的数据处理。但决策数算法也存在一定的缺点，它很难基于多个变量组合发现其相互之间的规则（王慧中等，2011）。决策树的分析结果常以树形图或者规则集的形式呈现，规则集的提取流程如图1所示：

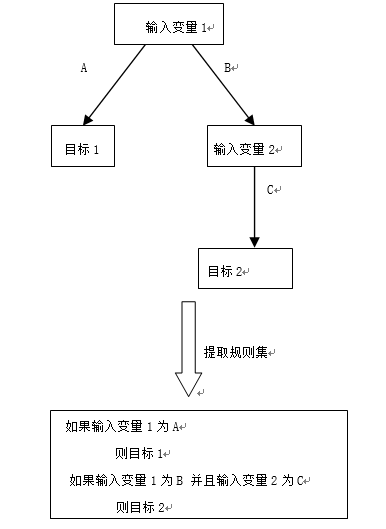


图1 规则集的提取流程

Fig1. Extraction process of rule set

1.3.2 随机森林算法

2001年 Breiman将自己提出的Bagging集成学习理论（1996）与Ho提出的随机子空间方法（1998）相结合，以多个决策树为基本分类器集合为一个学习模型，首次提出了随机森林算法（Breiman，2001）。随机森林是以多个决策树为基本分类器汇集成的一个集成学习模型，其中的决策树都是经过Bagging集成学习技术训练得到的。当输入需要分类的样本时，其中的决策树一起对数据进行分析，最终得到的分类结果由所有决策树分析出的最多的一种结果决定（董师师等， 2013）。其分类运行过程如图2所示。

随机森林算法作为一种比较新的机器学习算法，在运算量没有显著提高的情况下实现了更高的预测精度。随机森林算法能够有效地克服决策树的过拟合问题，且对多重共线性不敏感，结果对缺失数据和非平衡的数据也相对比较稳

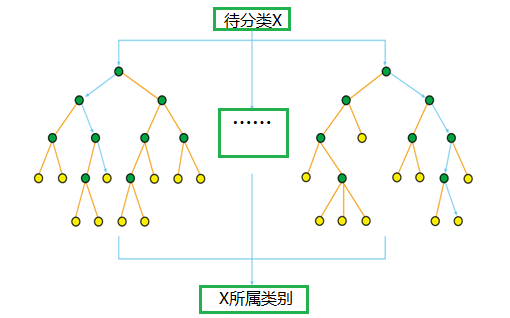


图2 随机森林的分类过程

Fig1. The classification process of Random Forest

健（李欣海，2013）。同时随机森林算法对噪声和异常值都有着比较不错的容忍性，且对高维数据分类的相关问题具有不错的可扩展性和并行性。除此之外，随机森林还是由一种数据驱动的非参数分类算法，所以只需通过对给定样本的学习训练相关的分类规则，不需要分类的先验知识。随机森林依靠其自身优秀的分析能力，自从研究出之后就被广泛地应用在各行业各领域中，在解决现实问题的过程中也取得了优秀的结果，受到了国内外学者的一致赞誉（董师师等，2013）。

1.4 灰飞虱的监测预警

对于危害农作物的主要病虫害进行监测预警是植物保护工作中重要的一部分，对于保障国家粮食安全和进行“统防统治”以及进行有效地控制是具有现实意义的。通过监测预警可以对即将爆发的病虫害进行提前预警预报，从而能够协助植保部门提前进行相关防控措施的准备，从而将危害控制在防治阈值的要求范围之内，进而服务于政府决策和农业生产。

灰飞虱*Laodelphax striatellus*（Fallén）活动区域十分广泛，主要位于菲律宾至西伯利亚一带的亚洲稻区和欧洲的温带地区（Kisimoto R，1989），它对相关地区的粮食生产带来了极大的负面影响,不仅吸食作物的汁液，且能够传播多种病毒，对农作物的危害极大，所以关于灰飞虱的生物特性的相关研究在国内外早已进行开展。

能够对灰飞虱的发生程度产生影响的气象因子主要包括：降水量、极大风速、平均气压、平均风速、平均气温、平均水汽压、平均相对湿度、日照时数、最大风速。在以往的传统研究中，一般认为对灰飞虱发生程度产生主要影响的气象因子是温度、湿度，另外包括作物品种、人类干预、天敌影响、耕作方式等其他环境因子。几种环境因子综合作用，共同产生了田间灰飞虱发生的最终结果。

1.4.1灰飞虱监测预警的重要性

灰飞虱属半翅目，飞虱科，灰飞虱属。灰飞虱的主要生活范围包括菲律宾至西伯利亚一带的亚洲稻区和欧洲的温带地区，在我国主要集中发生在长江流域及北部稻区。灰飞虱对于农作物的危害方式不是单一的，它的成、若虫能够利用刺吸式口器刺吸取食多种禾本科农作物、植物如水稻、小麦、大麦、李氏禾和双穗雀稗等植物，近年来，其对于玉米的危害也呈现正成逐步上升的趋势。

灰飞虱在通常喜欢田间通透性良好的环境，一般情况下灰飞虱栖息于植物植株位置较高的部位，并常向田边成群结队移动，因此，田边虫量相较于田内更多。灰飞虱的成虫翅型变化较稳定，越冬代的灰飞虱成虫以短翅型居多，其余各代以长翅型居多，雄虫除越冬外，其余各代几乎均为长翅型成虫。灰飞虱的成虫喜在植株嫩绿、高大茂密、作物生长条件好的地块产卵。雌虫产卵量一般为数十粒，越冬代最多，每头雌虫平均产卵量最高可达500粒左右，每个卵块的卵粒数，由1 - 2粒至10余粒，大多为5 - 6粒，种群生殖繁衍十分迅速。

灰飞虱在苏南地区一般一年发生6代，而在北方地区1年发生4 - 5代。在华北地区，越冬的灰飞虱若虫在4月中旬至5月中旬进行羽化，然后向农田迁移并产卵繁殖；第1代若虫于5月中旬至6月开始大量孵化，田间观测数目开始显著增加，1代若虫于5月下旬至6月中旬左右进行羽化；第2代若虫孵化于6月中旬至7月中旬之间，并于6月下旬至7月下旬羽化为成虫；第3代于7月至8月上、中两旬进行羽化；其第4代若虫于8月中旬至11月孵化，并于9月上旬至10月上旬进行羽化，有部分4代若虫则以3、4龄若虫进入越冬状态；其第5代若虫孵化于10月上旬至11月下旬左右，并进入越冬期。在全年发生中，以9月初的第4代若虫密度最大、数量最多。北方大部分地区多以第3、4龄和少量第5龄若虫在田边、沟边的杂草中越冬（丁锦华等，2002）。

当田间灰飞虱虫口较大时，作物的汁液将被灰飞虱大量吸食从而导致叶片枯黄，同时灰飞虱还会产生蜜露洒落在农田内及农田附近从而孳生大量霉菌，但较少出现类似褐飞虱和白背飞虱的“虱烧”、“冒穿”等症状。除了以成、若虫刺吸农作物进行危害以外，灰飞虱还可以携带病毒传播多种病害，如水稻黑条矮缩病、水稻条纹叶枯病、小麦丛矮病和玉米矮缩病等（王瑞等，2007）。研究表明，其传毒对农作物造成的实际危害远大于其成、若虫直接刺吸造成的危害（丁锦华等，2002）。2000年以来，由灰飞虱传播的水稻条纹叶枯病在江淮稻区发生日益严重，2004、2005年连续暴发成灾，造成严重减产甚至绝收。据不完全统计，2007年灰飞虱对山东地区粮食的危害面积达到63.6万 hm2；其中，对玉米的危害面积达到62万hm2，对水稻的危害面积达1.33 hm2以上，济宁、菏泽、泰安、枣庄、临沂等山东南部各市县受害较重，受害面积均达到了 4 ~ 12.5万hm2，虫口密度高达200 ~ 300头/m2，重发田甚至达到2000 ~ 3000头/m2（徐艳博等，2013）。

灰飞虱对于农作物的生长发育产生极大的威胁，所以建立灰飞虱的监测预警模型不仅有助于在植保工作中实施为更合理、更为有效的防止防控策略，还可以探究不同地区的不同环境因子对其发生程度的影响，从而分析灰飞虱爆发背后更为深层次的原因。

1.4.2 温度对灰飞虱发生程度的影响

温度是影响昆虫也是影响灰飞虱生长发育的重要因子（罗智心等，2009），不同温度对灰飞虱发育历期、存活率、翅型分化、雌雄性比等方面的都会产生一定的影响。在18 ~ 28℃范围内（张晓燕等，2014）。

灰飞虱属于温带地区的害虫，对于温带环境气候较为适宜，所以耐低温能力较强，而对高温的适应性较差，其生长发育的最佳适宜温度在28℃左右，冬季的低温对灰飞虱越冬若虫影响并不是很大，甚至在辽宁盘锦地区亦能够安全越冬，一般不会出现大量死亡的情况。灰飞虱要在低于 - 3℃且环境中持续时间较长时才会产生麻痹冻倒的现象，但除部分个体致死外，其余大部分个体仍能复苏。昆虫低温致死的重要原因是过低温度致使体液结冰，昆虫的体液能忍受0℃以下的低温而仍然保持不结冰，随着环境温度的进一步下降，当到达一定低温时，昆虫体温则会开始结冰，与此同时同时释放出热量，从而使体温复升；但当环境温度继续下降到达一定限度时，则虫体结冰，这个过程叫做昆虫体液的“过冷却现象”。通常来说，昆虫耐寒性的强弱和过冷却现象的关系是比较密切．即过冷却点越低．则耐寒性越强。因此，昆虫的耐寒能力强弱，经常可用它的过冷却点来表示，灰飞虱以3、4龄若虫越冬，过冷却点分别达到 - 7.22℃、 - 7.80℃（孙兴全等，2000）。当环境气温超过2℃且天气无风、晴天时，又能够爬至寄主植物的茎叶部进行取食并继续发育。

在18 ~ 28℃范围内，灰飞虱均能够完成各个生命历程的生长发育，且若虫发育历期随着温度的上升而缩短，在这其中18℃若虫发育最慢，28℃发育最快。然而当温度继续上升直至达到33℃时，总体来看若虫发育历期与28℃时的发育历期相比天数有所增加，且此温度下虫体发育缓慢，由此可见高温不利于灰飞虱的生长发育。关于灰飞虱的产卵量，一方面，产卵量的大小与雌虫的长短翅型有关，一般情况下短翅型雌虫产卵量要大于长翅型雌虫，除此外，温度条件对产卵量也有很大的影响，当温度升高，灰飞虱产卵量会相应增加，但当温度高于灰飞虱的适宜温度后，其产卵量反而会随温度升高而下降。研究表明在平均温度 18.8 ~ 28.5℃范围内（蒲茂华，1963），灰飞虱若虫发育速率随温度升高而加快；温度超过29℃，高龄若虫的速率就会降低，出现滞育和死亡现象，高温会使灰飞虱进入越夏滞育阶段。此时灰飞虱停止取食，同时这个阶段内若虫死亡率较高，且不能完成羽化（高东明等，1994；刘向东等，2007）。根据研究统计，在一定的温度范围之内，温度和灰飞虱的生长发育程度表现出类似于抛物线的关系。

1.4.3湿度对灰飞虱发生程度的影响

与降水量、气压、水汽压、平均相对湿度、日照时数等相关的气象因子均能够对空气及土壤中的湿度产生影响，进而影响到灰飞虱的生长发育各个阶段。湿度对灰飞虱的影响主要体现在浸水后对其卵的致死率上，而相对于卵具有不同的浸耐力，灰飞虱的成虫相对相对更喜欢潮湿的环境，因为潮湿环境不仅更适合灰飞虱其他生命活动，也能够为灰飞虱提供更多的食物。

灰飞虱喜欢生长在湿度比较高的环境中，其卵的短期耐浸能力比较强，在卵的发育前期、中期，浸泡1d的死亡率分别为0 % 和7 % ，但随着浸泡时间的增加，在第六天其卵发育的前期、中期、后期的死亡率也逐步增加到62 % 、68 % 、100 % 。而灰飞虱在卵发育的后期，耐浸泡能力显著降低，浸泡1d后卵的死亡率即可达到45.3 % 。

1.4.4 风速对灰飞虱的影响

风速会对具有迁飞性昆虫的迁飞活动提供助力，迁飞性昆虫借助气流的运载，可以实现中场距离的种群迁移。

关于灰飞虱的虫源，直到1998年，国内还一直普遍认为是主要虫源都是本地虫源（李济宸等，1998），但是在国外早已有关于灰飞虱远距离迁飞的研究报道：最早可追溯到1968年日本的Asahina和Tsuruoka报道在中国东海的气象观测船上捕到了一些灰飞虱。后来，又报道了分别在1969年、1977年、1978年和1979年在中国东海气象观测船上，分别捕到灰飞虱96头、734头、367头和1025头；福建省农科院植保所于1978年、1979年也在东海捕到了灰飞虱，其数量分别为50头和516头（刘浩官等，1983）；之后Otuka又论证了日本西部带病毒的灰飞虱是从我国的江苏省跨海迁飞过去的（Otuka，2010）。明确灰飞虱的迁飞特性，对灰飞虱监测预警和源头治理及减少其对粮食作物的危害具有至关重要的意义（张海燕等，2011）。

具有迁飞性的昆虫飞过边界层后主要借助于上空水平气流的运载而迁飞到远处，其迁飞方向和速度都和当时上空的风向，风速较为一致。我国东半部春、夏季，由于太平洋副高压的逐步向北推进，高空经常盛行较为强劲的偏南风（张孝羲，1980），所以盛行偏南风为安徽地区与江苏地区的灰飞虱提供了极为便利的运载气流，帮助他们向北迁移至山东南部地区、半岛地区，甚至更为向北。

1.4.5国内灰飞虱已建立的监测预报模型

目前，关于灰飞虱发生程度的不同预测模型已经开始建立并取得一定进展，但多数监测预警模型探索的都是线性关系，且较多关注大范围区域的灰飞虱发生规律，而针对特定地区灰飞虱发生规律的非单纯线性监测预警模型的建立尚在研究探索阶段。

汪恩国利用临海市1986 ~ 2005年灰飞虱成虫灯下诱集的资料和田间测报调查情况，建立了关于临海市灰飞虱种群数量预测的7种数学模型，分析总结了灰飞虱种群变化特征及其运动规律，并阐述了灰飞虱的田间消长动态，对建立灰飞虱的监测预警模型进行了尝试，但模型的回测检验吻合率较低，仅有54.69 % （汪恩国，2007）。

顾国华等针对一代灰飞虱的不同虫态发生期,构建了四个不同温度条件下的虫态历期模型，分别为：成卵前期模型、卵历期模型、若虫历期模型。同时，将通州市气象站提供的1987 ~ 2006近20年来3月中旬至6月上旬的平均气温，利用统计方法划分成5种类型：高温、偏高、正常、偏低、低温，以某气温类型某虫态发生期内的历史平均气温作为输入变量进行预测分析，经过对2006和2007年田间实际虫情进行验证和试预测，取得了非常理想的效果，吻合率和准确率均达到100 % 。该研究根本目的在于建立合理科学的监测预警模型以便于在害虫防治中的应用，了解一代灰飞虱1龄若虫发生期可以为适时开展一代灰飞虱的田间防治提供帮助，从而能够达到压低田间虫量基数、减少一代成虫数量的效果，进而能够减轻由灰飞虱传毒造成的病毒病的危害（顾国华等，2008）。

胡英华等利用2005 ~ 2011年室内饲养及田间调查记录的济宁地区灰飞虱发生量数据，明确了济宁地区灰飞虱常发生5代，并通过测报记录资料以及济宁地区气象资料，在DPS平台中采取逐步回归分析方法，组建了灰飞虱主要为害世代第1代的发生量及发生期的线性预测模型，取得了较好的效果，回测检验拟合准确率较高，达到97 %（胡英华等，2011）。

包云轩等（2014）利用1979~2011年NCEP（美国国家环境预测中心）提供的逐日气象再分析资料，以及长江中下游稻区的白背飞虱逐候灯诱数据，通过对白背飞虱迁入量和主要的大气环流特征量进行相关性分析，建立了涉及长江中下游稻区13个植保站的白背飞虱迁入始见期、北迁高峰期、南迁高峰期和终见期4个时间阶段的关于白背飞虱候发生程度BP神经网络短期预测预报模型。

其中选取了与白背飞虱迁入量相关显著的大气环流特征量为预报因子，按5级发生程度对白背飞虱迁人量进行分级处理，建立了迁入始见期、北迁高峰期、南迁高峰期和终见期白背飞虱候发生程度共4个BP神经网络预报模型，模型的预先检验取得了较好的结果，准确率稳定在80 % 以上，可应用于相关地区的白背飞虱短期预测预报。研究结果一方面对于揭示气象因子对白背飞虱迁入和发生的影响规律，作好其发生程度的预测预报，适时、有效防控白背飞虱为害具有积极意义；另一方面，为我们建立灰飞虱的监测预警模型提供了参考借鉴和思路指导。

1.5 立题依据与意义

山东省一直以来都是我国重要的产粮大省之一，据国家统计局发布的调查公告显示，2016年全年，山东粮食种植面积为7511.5 khm2，单位面积产量达到6258.0 kg / hm2，总产量为47007 kt。所以山东粮食生产的安全直接影响到我国整体粮食安全以及人们的生活、生存保障，具有重大意义。而近几年来，灰飞虱逐渐成为山东省主要虫害之一，灰飞虱可大面积的传播多种病毒，其直接危害及传毒危害的作物如：水稻、小麦、玉米等，在山东省都是主要粮食作物，有大面积种植，是山东省粮食生产的主要品种。

灰飞虱的抗药性和环境适应性等能力都较强，另外再加之近年来耕作制度过于粗放、轻简栽培技术在粮食生产中被大面积应用，以及粮食品种较多、较乱、较杂等各种原因，最终形成一种灰飞虱种群发生较为适宜的农田生态环境。这种情况直接导致了灰飞虱种群暴发危害呈现持续性的特点，特别是循环不断的寄主植物链更成为灰飞虱及其传播的病毒病发生发展的主要促进因素（孙广仲等，2006）。2008 年，由于较适宜的温湿度气候条件，济宁地区灰飞虱在5月25日 ~ 6月10日进入发生高峰期，发生数量巨大、持续时间较长，且灰飞虱的带毒率高达45% ，从而致使济宁市的12.7万 hm2 麦套玉米田和早直播玉米田全部受害，大部分重播改种或绝收，直接经济损失达 9 亿元（苏加岱等，2008；苏加岱等，2009）。

灰飞虱对山东省的粮食安全产生了较大的威胁，且常常呈短期内大爆发的情况。为了控制灰飞虱的发生，通常采用化学药剂进行防治，但是这存在一定的局限性，一方面由于表面喷施玉米叶片不能够完全地隔断传毒灰飞虱的迁飞和危害程度，起不到彻底杀灭的作用；另一方面，由于长期大量、不合理地施用化学药剂等杀虫剂，造成灰飞虱对多种杀虫剂表现了出不同程度的抗药性，杀灭效果也逐渐减退（刘向东等，2006；Zhu Jinliang *et al*.，2009；王彦华等，2010；孔晓明等，2013）。为了防治灰飞虱，所以只好继续施用农药，这种恶性循环一方面导致了灰飞虱抗药性的进一步增强，防治效果将逐年下降，另一方面使得粮食中的农药残留量上升，影响了食品的品质与安全，所以建立灰飞虱的监测预警模型，探究影响灰飞虱发生的主要因素，以及通过预测结果合理把控农药的施用量，具有重要意义。

目前，国内外虽然已经有关于灰飞虱监测预警模型的构建，但绝大部分研究都存在一定的局限。要么在研究中仅仅采用有限的气象因子，如仅使用温度及湿度、温度及光照等；要么仅采用单纯的线性回归分析，无法探究不同影响因子之间的相互作用，准确率不够理想；要么探究的都是大面积区域内的、长时间尺度的灰飞虱发生规律。而针对于特定地区灰飞虱发生程度的非单纯线性、高准确率的监测预警模型还未建立，同时也未能从绿色防控的角度对灰飞虱的防治工作提供参考意见。

因此，以通过对济宁地区多年灰飞虱发生量以及多种气象因子的统计结果为数据支撑，以决策树算法及随机森林算法为主要数据分析方法，利用R语言为数据处理平台，从而建立一种关于济宁地区灰飞虱发生量的科学有效的监测预警模型，来有效地发挥大数据技术在植物保护中的应用价值，为灰飞虱的科学绿色防控提供服务，为相关植保工作的开展提供参考。

## 2 材料与方法

2.1 数据来源

2.1.1 气象数据来源

2008年 ~ 2016年的气象数据来源于由国家气象信息中心和中国气象局气象数据中心维护管理的“中国气象数据网”（http://data.cma.cn/），其中的逐日气象观测资料为实验所需气象数据。

2.1.2 灰飞虱数据来源

济宁地区灰飞虱发生程度2008 ~ 2013年的资料主要来源于济宁植保站在山东省济宁市任城区对的田间调查；2014 ~ 2016年的资料主要来源于我们自己的田间调查，调查地点同样位于济宁市任城区。

调查时小麦品种为济麦22，种子包衣，地块类型为中高肥水田，种植方式为“小麦 - 玉米轮作”。选择生长均匀一致，有代表性的麦田5块作为项目观测田进行调查取平均值，每块麦田面积不少于667m2，田块间距不小于2km。调查时间保持连续8年，自2008年至2016年，每年调查时间范围主要选择在小麦苗期、返青拔节期至小麦收获止，每次调查至少间隔7天，每年调查次数不少于7次。 调查方法为盘拍法，按平行多点跳跃法取点调查并换算为虫口密度（万头/667m2）。

2.2 灰飞虱发生程度分级及因变量分级

根据《麦田灰飞虱发生程度分级标准》，将济宁市灰飞虱发生程度以每亩虫口密度为分级依据，将经过调查得到的灰飞虱发生等级分为5级，分别为轻发生（1级），中等偏轻（2级），中等发生（3级），中等偏重（4级），大发生（5级），如表1所示。另外，在实际的植保工作中，通常根据虫害发生程度的不同选择不同的防治方法，因此我们因变量Y按照发生程度划分为“slight”和“heavy”。将1级、2级发生级别较小时命名为“slight” ，此时采用物理防治或生物防治的措施；将3级、4级、5级发生级别较大时命名为“heavy”， 可采用化学方式为主、其他防治方式为辅的措施 ，从而提高监测预警模型的准确度，为植保工作提供绿色防控策略与参考。

表1 麦田灰飞虱发生程度分级标准

Table 1 Grading standard of occurrence degree of *Laodelphax striatellus* in wheat field

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 发生级别  Occurrence level | 轻发生  （1级）  Slight  occurrence  (Level 1) | 中等偏轻  （2级）  Moderate  slight  (Level 2) | 中等发生  （3级）  Moderate occurrence  (Level 3) | 中等偏重  （4级）  Moderate emphasis  (Level 4) | 大发生  （5级）  Heavy  occurrence  （Level 5) |
| 虫口密度  （万头/667m2）  Insect density  （ten thousand/667m2） | ＜2 | ≥2，＜ 10 | ≥10，＜30 | ≥30，＜50 | ≥50 |

2.3 自变量的选择

一方面，不同的温度、土壤湿度、33℃以上和 - 8℃以下的极端温度等对于灰飞虱的存活、产卵、卵孵化、越冬等生理行为都有着重要影响（孙兴全等，2000；胡英华等，2010；张晓燕等，2014）。另一方面，灰飞虱具有远距离迁飞的特性，所以济宁地区虫害来源有本地越冬和外地迁入两种（张海燕等，2011），而气压、风速等气象条件的变化都将对灰飞虱的迁飞造成影响。所以选择气象因子“平均日降水量”、“平均气压”、“平均气压日较差”、“平均风速”、“平均最大风速”、“平均温度”、“平均温度日较差”、“平均水汽压”、“平均相对湿度”、“平均20 - 20时日照时数”、“是否有33℃以上及 - 8℃以下极端温度天气”，共计11个。为了探究不同时间尺度的气象因子对灰飞虱发生程度的产生的短期中期长期影响，根据张爱民等对成虫寿命和繁殖力的研究（张爱民等，2008），选择“30天”、“60天”、“90天”三种时间尺度，监测预警模型中全部的气象因子总数为33个。

2.4 随机森林算法运行原理

之前在前文中提到过，随机森林的是以多个决策树为基本分类器汇集成的一个集成学习模型。集成学习可行有两个前提条件：一是单个基本分类器是有效的，也就是说单个分类器的精度应该大于随机猜对的概率；二是各个基本分类器是有差异的，要达到差异性，可以通过采用不同的训练样本或者不同的训练方法实现。

随机森林作为一种集成学习模型，也需要满足以上两个条件。所以随机森林中的决策树都是经过Bagging集成学习技术训练得到的，而对于这些决策树的选择则依照袋外估计理论进行。随机森林依据这两方面，保证了每个分类器的正确率与基本分类器的差异性（董师师等，2013）。所以随机森林的主要原理分为三个部分，分别是决策树算法原理、Bagging集成学习算法和袋外估计理论（Breiman L，2001）。

2.4.1 决策树算法运行原理

当今决策树算法中，较为成熟的是C5.0算法，它同其它决策树算法一样，是以树形图形式显示最终结果的。决策树C5.0算法共涉及3个函数，分别是计算熵值函数、计算信息增益函数和计算信息增益率函数。其中，熵值函数是决策树的变量选择函数，用来预测信息位数。关于熵值函数的计算公式如下：

Entro(p1, p2, …, pn) = - p1log2(p1) - … - pnlog2(pn),其中pn为n发生的概率

当Entro(p1, p2, …, pn) = 0，表示存在唯一的可能性；pn的差别越小，Entro(p1, p2, …, pn)的值就越大，相反pn的差别越大，熵值就越小。

决策树中信息熵的计算公式如下：

其中，m是一个样本集合，目标变量n有k个，freq(ni, m)表示n的样本数，｜m｜表示集合m的样本数。

根据计算所得的信息熵值计算信息增益值，信息增益函数是进行变量选择前后的信息差值的函数。S是某属性变量，有a个分类，其计算公式如下：

Info(T)和Info(S)分别是决策树进行属性划分前后的信息增益值，其计算公式如下：

信息增益率则是逐个计算比较，最终确定树形图上节点的位置。

C5.0算法模型的置信度是统计预测值中正确值的个数占总样本数的比值。置信度高说明该模型的预测结果较准确。

2.4.2 Bagging集成学习算法

Bagging集成学习算法可以从原始样本集中进行可重复抽样，从而得到不同的Bootstrap训练样本，进而训练出各个基本分类器，即各个单个决策树，其过程如图3所示。假设原始样本集的样本容量大小为N，每次有放回抽取的Bootstrap样本大小也为N，那么每个样本未被抽中的概率约为(1 - 1/N )N，当N很大时，这个概率值则 ≈ 0.368。这说明每次抽样时原始样本集中约有37 % 的样本从未被抽中，这一部分未被抽中的样本即被称为袋外数据，而另一部分被抽取的样本则是用于构建模型的Bootstrap训练样本。

Breiman指出对于决策树等不稳定（即对训练数据较敏感）的分类器，Bagging集成学习算法能有利于其提高分类的准确度。此外Bagging算法可以并行训练多个基本分类器，可以节省大量的时间开销，这也是该算法的另一个重要优势。

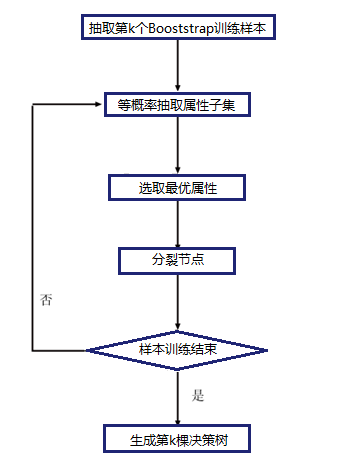


图3 随机森林中单个决策树选择的过程

Fig3. The process of selecting a single decision tree in Random Forest

2.4.3 袋外估计理论

前文提到过，由于Bagging集成学习算法每次从原始本集中随机抽取Bootstrap训练样本时，约有37 % 的样本最终没有被选中，这一部分未被选中的样本即被成为袋外数据。它们可用于估计随机森林的单棵决策树分类强度以及决策树之间的相关性，进而可得到随机森林泛化误差界的估计，这个过程就叫做袋外估计。Breiman在论文中指出袋外估计是一种无偏估计，袋外估计与用同训练集一样大小的测试集进行估计的精度是一样的。

在随机森林模型运行的过程中，Bootstrap训练样本会被用来进行模型的构建，而袋外数据则被用来对构建的模型进行检验，当袋外估计错误率（OOB estimate of error rate）较低时，则说明随机森林的准确度较高，从而具有科学性与实用性。

3 结果

3.1 数据的整理导入及调用

本文的数据处理使用RStudio实现。首先，将33个自变量命名为x1，x2，x3,……x33，如表2所示。其次，将整理好的71组数据在EXCEL表格中保存为csv格式，命名为“Laodelphax striatellus.csv”，导入R平台，读取为数据集“data.Laodelphax”。然后，安装randomForest包，并使用library指令进行调用，设置编号为“518”随机森林种子，保证实验结果的可重复性。最后，将71组数据按7 : 3的比例划分为训练数据集“traindata”和测试数据集“textdata”,其中以训练数据集建立监测预警模型，以测试数据集对模型进行检验。

代码如下：

data.Laodelphax<-read.csv("C:/Users/Administrator/Desktop/Laodelphax striatellus.csv",header=T) # 读取原始样本，命名为“data.Laodelphax”

install.packages("randomForest") # 安装randomForest程序包

library(randomForest) # 调用randomForest程序包

set.seed(518) # 选择编号为518的随机森林种子

ind<-sample(2, nrow(data.Laodelphax), replace=TRUE, prob=c(0.7, 0.3))

traindata<-data.Laodelphax[ind==1,]

testdata<-data.Laodelphax[ind==2,] # 将原始样本按照7：3的比例划分为训练数据和测试数据

值得注意的是，这其中有一步与其余大部分文章中使用的随机森林代码不同的地方，即“set.seed(518) # 选择编号为518的随机森林种子”这一步。之所以选择某编号的种子是因为随机森林的结果输出是一个不确定的随机结果，如果缺少相应的前提设定，那么随机森林每次运行的各结果，例如：mtry取值错误率（errRate）、基尼系数（MeanDecreaseGini）、袋外估计错误率（OOB estimate of error rate）等结果虽然总体大致相同，但都会有一些细微的偏差。这一步的意义在于，确定一个固定的结果，以保证实验的可复制性。

另外，使用ind指令将原始样本按照7：3的比例划分为traindata（训练数据）和testdata（测试数据）的原因，是为了进行二次检验来避免随机森林自身出现极小概率的错误，从而保证结果的准确性。

表 2 自变量命名对照

Table 2 The contrast of independent variable’s alternate name

|  |  |  |  |
| --- | --- | --- | --- |
| 气象因子  Meteorological factor | 时间尺度30天  time scale 30 days | 时间尺度60天  time scale 60 days | 时间尺度90天  time scale 90 days |
| 平均日降水量（0.1mm）  Average daily precipitation | x1 | x12 | x23 |
| 平均气压（0.1hPa）  Average atmospheric pressure | x2 | x13 | x24 |
| 平均风速（0.1m/s）  Average wind speed | x3 | x14 | x25 |
| 平均温度（0.1℃）  Average temperature | x4 | x15 | x26 |
| 平均水汽压（0.1hPa）  Average water vapor pressure | x5 | x16 | x27 |
| 平均相对湿度（RH%）  Average relative humidity | x6 | x17 | x28 |
| 平均20-20时日照时数（0.1h）  Average sunshine hours from 20 to 20 | x7 | x18 | x29 |
| 平均气压日较差（0.1hPa）  Daily range of Average atmospheric pressure | x8 | x19 | x30 |
| 平均最大风速（0.1m/s）  Average maximum wind speed | x9 | x20 | x31 |
| 平均气温日较差（0.1℃）  Daily range of average temperature | x10 | x21 | x32 |
| 时间尺度内有无极端温度天气（是/否）  Is there any extreme temperature weather in the time scale | x11 | x22 | x33 |

3.2 mtry值的确定

mtry值的大小意味着随机森林中在每个决策树节点上用于进行分类的气象因子的个数，而对于不同的随机森林模型应采用不同的mtry值，以降低模型错误率（赵北庚，2015）。根据随机森林分析得出最佳mtry值为16，此时对应的错误率仅有0.1555，如表3所示。将33个气象因子作为自变量X，将发生程度severity作为因变量Y，用于构建测试模型1，代码如下：

n<-ncol(traindata)-1 # 以训练数据为样本，确定循环次数

for(i in 1:n){set.seed(518) # 遍历所有mtry的可能取值，选取编号为518的随机森林种子

test\_model\_1<randomForest(severity ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+

x11+x12+x13+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26+

x27+x28+x29+x30+x31+x32+x33,data=traindata,mtry=i, importance = TRUE,

proximity= TRUE) # 建立测试模型1

errRate<-mean(test\_model\_1$err.rate)

print(errRate)} # 输出mtry各种取值对应的错误率，从中选取错误率最小的结果，作为决策树节点所选择的最佳影响因子个数。

表3 全部33个mtry值对应的错误率

Table 3 Corresponding error rate of all 33 mtry values

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| mtry值  mtry values | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 错误率  Error rate | 0.243 | 0.207 | 0.213 | 0.226 | 0.198 | 0.231 | 0.184 | 0.211 | 0.181 | 0.169 | 0.172 |
| mtry值  mtry values | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 错误率  Error rate | 0.195 | 0.173 | 0.170 | 0.178 | 0.155 | 0.188 | 0.168 | 0.164 | 0.159 | 0.168 | 0.184 |
| mtry值  mtry values | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| 错误率  Error rate | 0.176 | 0.176 | 0.159 | 0.167 | 0.178 | 0.179 | 0.177 | 0.171 | 0.183 | 0.181 | 0.174 |

3.3 ntree值的确定

ntree值的大小代表了随机森林模型中使用的决策树的数量，虽然通常来讲ntree值越大，模型的准确率越高，但在模型的实际构建中，ntree值达到一定大小时，模型的准确率就不再提高了，所以选取合适的ntree值是对模型的优化，避免模型过于复杂。根据随机森林分析得出当ntree值为100时，准确率已不再增加，故选择100最作为最佳ntree值，如图4所示。

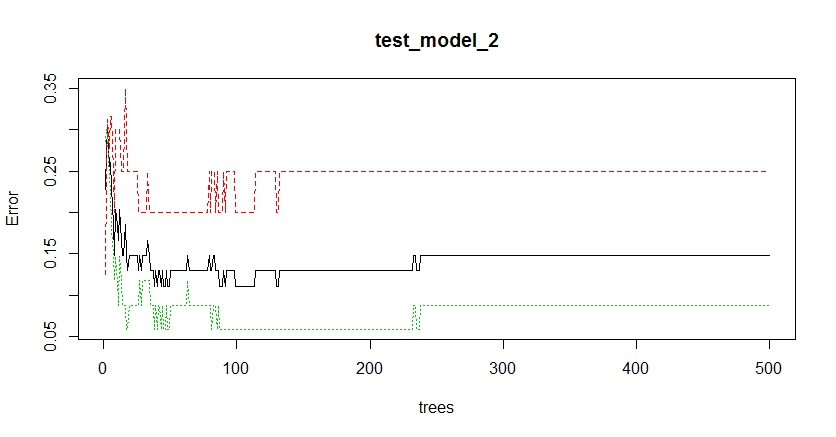


图4 ntree值与随机森林错误率的关系

Fig4. The relationship between error rate of Random Forest and ntree values

构建测试模型2，代码如下：

set.seed(518)# 选取编号为518的随机森林种子

test\_model\_2<randomForest(severity~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+

x11+x12+x13+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26+x27+x28+x29+x30+x31+x32+x33, data=traindata,mtry=16,importance=TRUE,

proximity=TRUE) # 建立测试模型2

plot(test\_model\_2) # 绘制树的数量与错误率之间的关系图，以此选取错误率最先达到最小的结果，作为最佳ntree值

3.4构建监测预警模型并分析主要影响因子

一方面，将最佳mtry值16、最佳ntree值100，用于构建正式的监测预警模型，并输出模型结果及袋外估计错误率，如表4、表5所示；另一方面，使用importance及VarImpPlot指令（李欣海，2010），显示33个气象因子的基尼系数（MeanDecreaseGini），并绘制图形排序，如图2所示。经模型分析，袋外估计错误率为11.11 % ， x20、x31为主要影响因子，x12、x1为重要影响因子。

代码如下：

set.seed(518) # 选取编号为518的随机森林种子

result.rf<randomForest(severity~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+

x12+x13+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26+x27+x28+x29+x30+x31+x32+x33,data=traindata,mtry=16,ntree=100,importance=TRUE,

proximity=TRUE) # 建立随机森林模型并分析数据

result.rf # 输出随即森林结果

varImpPlot(result.rf,type=2) # 用图形表示每个自变量的基尼指数

表4 随机森林袋外估计误差率

Table 4 OBB estimate of error rate in Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
|  | heavy | slight | class.error |
| heavy | 16 | 4 | 0.2000 |
| slight | 2 | 32 | 0.0588 |
| total | - | - | 0.1111 |

表5 每个自变量的基尼系数（MeanDecreaseGini）

Table 5 MeanDecreaseGini of each independent variable

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 气象因子  Meteorological factor | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | x10 | x11 |
| 基尼指数  MeanDecreaseGini | 1.059 | 0.739 | 0.144 | 0.358 | 0.522 | 0.254 | 0.690 | 0.222 | 0.805 | 0.276 | 0.000 |
| 气象因子  Meteorological factor | x12 | x13 | x14 | x15 | x16 | x17 | x18 | x19 | x20 | x21 | x22 |
| 基尼指数  MeanDecreaseGini | 1.912 | 0.222 | 0.464 | 0.208 | 0.136 | 0.162 | 0.363 | 0.718 | 6.305 | 0.126 | 0.000 |
| 气象因子  Meteorological factor | x23 | x24 | x25 | x26 | x27 | x28 | x29 | x30 | x31 | x32 | x33 |
| 基尼指数  MeanDecreaseGini | 0.355 | 0.804 | 1.040 | 0.240 | 0.035 | 0.087 | 0.222 | 0.419 | 5.216 | 0.383 | 0.003 |

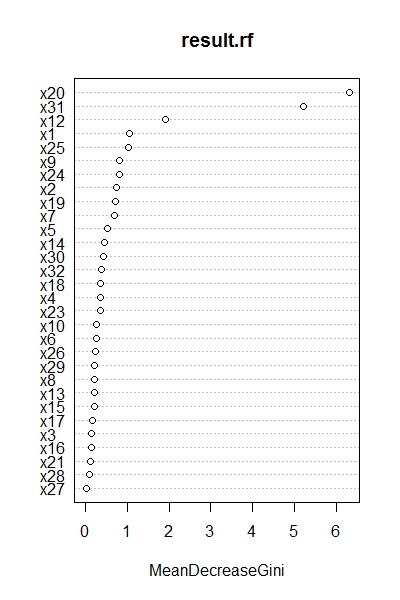


图5 对各种气象因子的基尼系数进行排序

Fig.5 Sort all kinds of meteorological factors by their MeanDecreaseGini values

根据随机森林模型的分析结果我们可以看出，在影响济宁地区灰飞虱发生程度的各气象因子中，“累计60天平均最大风速”与“累计90天平均最大风速”对济宁地区灰飞虱的发生起到了主要作用。另外，“累计60天平均日降水量”和“累计30天平均日降水量”对灰飞虱的发生起到了一定的作用。

使用partialPlot指令，分别绘图（李欣海，2010），表示这四个气象因子对灰飞虱发生程度的影响,如图3、4、5、6所示。代码如下：

partialPlot(result.rf,traindata,x20,"heavy",main="",xlab="累计60天平均最大风速（0.1m/s）",ylab="自变量影响程度")

partialPlot(result.rf,traindata,x31,"heavy",main="",xlab="累计90天平均最大风速（0.1m/s）",ylab="自变量影响程度")

partialPlot(result.rf,traindata,x12,"heavy",main="",xlab="累计60天平均日降水量（0.1mm）",ylab="自变量影响程度")

partialPlot(result.rf,traindata,x1,"heavy",main="",xlab="累计30天平均日降水量（0.1mm）",ylab="自变量影响程度")

3.5 模型准确率的检验

在使用训练数据建立模型之后，我们将使用测试数据检验模型的正确率。使用table指令，将textdata导入监测预警模型result.rf 中进行检验，得出模型正确率为88.24 % ，如表6所示。代码如下：

table(actual=testdata$severity,predicted=predict(result.rf,newdata = testdata,type = "class"))

# 利用测试数据检验模型的可信度，并记录输出结果

表6 测试数据检验错误率

Table 6 Test the error rate by test data

|  |  |  |  |
| --- | --- | --- | --- |
|  | heavy | slight | class.error |
| heavy | 9 | 0 | 0 |
| slight | 2 | 6 | 0.3333 |
| total | - | - | 0.1176 |

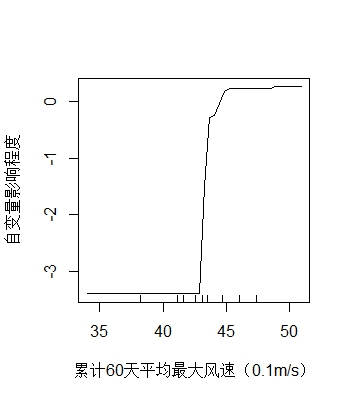


图6 累计60天最大风度对灰飞虱的影响

Fig.6 The influence of cumulative 60 days average maximum wind speed on *Laodelphax striatellus*

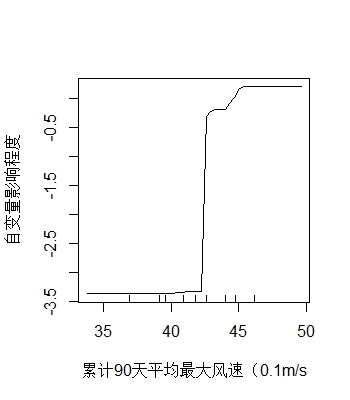


图7 90天最大风度对灰飞虱的影响

Fig.7 The influence of cumulative 90 days average maximum wind speed on *Laodelphax striatellus*

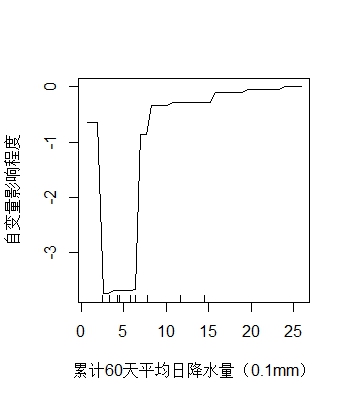


图8 60天平均日降水量对灰飞虱的影响

Fig.8 The influence of cumulative 60 days average daily precipitation on *Laodelphax striatellus*

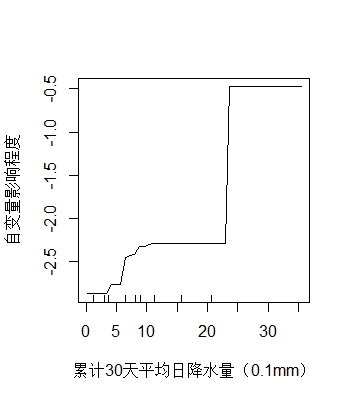


图9 30天平均日降水量对灰飞虱的影响

Fig.9 The influence of cumulative 30 days average daily precipitation on *Laodelphax striatellus*

3.6 使用监测预警模型进行预测

在实际应用及操作中，为了通过模型实现对灰飞虱发生程度的预测，可以通过以下方法进行，然后根据预测的结果决定采用何种程度的防治方法。

以将预测当天与预测前一天的33个气象因子按照标准格式进行整理并保存为“csv”格式的excel文件。这里使用2016年最后两次调查结果进行举例，这两次调查中因变量Y（灰飞虱发生程度），都为“slight”，首先因变量清空，然后使用table与 predict指令代入检验。

代码如下：

pre.data<-read.csv("C:/Users/Administrator/Desktop/predata.csv",header=T)

# 读取数据predata作为预测数据

table(actual=pre.data$severity, predicted=predict(result.rf,newdata = pre.data,type = "class")) #利用模型进行预测

此处采用两组数据举例说明，结果输出如下：

Predicted Result

actual heavy slight

slight 0 2

heavy 0 0

随机森林模型对着两组数据的分析结果显示，灰飞虱的发生程度为“slight”，原实际结果也为“slight”，预测结果与原结果相符合，建议此时采用程度较轻的防治措施以减少田间农药施用量。

## 4 讨论

## 4.1 风速与降水对灰飞虱发生程度的影响

从结果中可以看出，中、长期的累积风速对灰飞虱的发生起到了主要作用，分析其原因，主要是由于灰飞虱既是能当地越冬的本源性害虫，又是具有季节型远距离迁飞的媒介昆虫，山东南部周边各省份如安徽（王丽等，2011）、江苏（张海燕等，2011）等地均有灰飞虱迁入济宁地区，而风速的大小对其迁飞活动起到重要影响。一者，由于调查时间选择在当地小麦苗期、返青拔节期至小麦收获之间，而这个季节的气流主要为来自南方的暖空气，强劲的暖气流为灰飞虱的远距离迁飞提供了运载气流（张海燕等，2011），而迁飞性昆虫也会主动寻找和利用有利的季节性气流来实现远距离迁飞（G Hu和KS Lim等，2016），所以当风速不足时，则将不利于灰飞虱的迁飞行为；二者，这个季节很难有超出灰飞虱承受能力的极端大风天气，所以在调查数据中体现出风速越大灰飞虱发生程度越高的情况；三者，灰飞虱的迁飞是一个爆发性的过程，时间尺度内大风天气数越多则爆发风险越高，但由于调查时间从深冬季节开始，所以“90天”的长时间尺度内的大风天气不一定都是挂南风的，也有一些北风，这部分刮北风的大风天气对灰飞虱的迁飞是无法起到帮助的，所以“累计60天平均最大风速”这一气象因子造成的影响是大于“累计90天平均最大风速”的。

短、中期的降水对灰飞虱的发生起到了一定作用，这是因为灰飞虱喜欢生长在湿度比较高的环境中，其卵的耐浸能力比较强，虽然过高的湿度将对其卵发育的后期产生负面影响（孙兴全，2000），但也是由于调查期间并不是济宁地区的多雨季节，所以在这个时间段，更多的降水将带给灰飞虱更适宜的生长条件，相反，缺少降水则将会对灰飞虱的发生产生负面影响。而在时间尺度上，可能是由于较长时间之前的降水将会被蒸发，弱化了其起作用，所以其主要作用的降水相关气象因子为短中期。

由分析结果我们可以得出结论，在中长期有大风且短中期有降水的情况下，有较高的灰飞虱爆发风险，所以可以据此提前采取较大程度的防治措施，如提前喷洒农药等，从而避免灰飞虱的大爆发。

4.2 模型对于其他地区或其他昆虫的应用分析

对于济宁地区灰飞虱发生程度监测预警模型的构建流程及构建思路，为山东其他地区或其他虫害的监测预警模型的构建提供了参考和一套框架体系。构建过程中对涉及详细的方法及R语言代码都做了十分详细的记录，并进行的补充说明。如此一来在构建其他地区或其他虫害的监测预警模型时，可以按照已有的流程、方法、代码直接进行相关数据的填充，从而能够在较短的时间内构建出另一套以气象因子为主要自变量的针对不同对象的监测预警模型。

除此之外，在构建其他地区或其他虫害的监测预警模型时，如果能够得到新的自变量，如人为因素、生物因子、土壤状况因子等，只需要对自变量数据重新进行整理、补充、定义和命名，再应用进监测预警模型的构建中即可。

4.3 研究的优势及展望

根据前人的研究，随机森林算法是优于如逻辑回归、决策树、支持向量机、自适应助推法、神经网络等大部分算法的（Pal M，2005；Cutler等，2007；Chan等，2008）。得以于使用的算法优秀，本文中构建的灰飞虱监测预警模型达到了较为理想的效果。

目前对于随机森林在各个领域的应用研究都有着不错的进展，其在农业领域的应用也逐步开始崭露头角，如在小麦叶片SPAD值遥感估算（王丽爱等，2015）、玉米发育程度自动测量（石礼娟等，2017）、棉蚜虫害等级预测（李磊等，2017）等方面都展现出其应用价值。而不同于以往对于随即森林的研究，本文在构建灰飞虱的监测预警模型的过程中，针对随机森林算法在以下几个方面进行了突破：一是更加重视在随机森林在R语言的体现，列举详细的代码，以保证模型构建之后的易操作性；二是重视seed的设置以确保随机数产生的一致性，进而得以保证实验结果的可重复性；三是对模型的具体使用方法提供了指导，保证了模型的实用性。另外，在气象因子的选择上考虑了不同时间尺度对于结果的影响，从而构建了多种类多时间尺度的高准确率分类模型。

然而关于我们的模型，还是存在一定的不足之处，预测准确率并没有达到完美的效果，分析其原因，主要有以下几个方面。首先，由于历史数据的限制（杨波等，2014）,只能够获得2008年 ~ 2016年9年的灰飞虱发生数据，使得数据量稍小，进而为了保证正确率只能够将自变量划分为两类进行讨论，无法沿用之前5个类型的分级标准。其次，本模型仅考虑气象因子而未能考虑其他环境因子的影响，如一些生物及非生物因子等，使得模型具有一定的局限性。再次，灰飞虱作为一种具有迁飞性的害虫，仅仅针对为害地区进行研究调查是不够的，应该综合考虑虫源地区与为害地区的各种环境因子之间的关系，这需要多方面的配合和支持。最后，相关实验过程在R语言中的中代码还未达到最佳优化，造成该模型没有达到最简便、优秀的程度。

我们认为如果能够通过接下几年来灰飞虱数据、气象数据和其他环境因子数据的补齐，可以使模型更加精确。另外，灰飞虱迁飞性在结果中的体现也是一个值得进一步深入研究的方向，可以不仅仅局限于一个地区的研究，而把整个我国东部地区的各省分的发生情况综合考虑，建设一个全面的监测预警系统。相信在今后通过农业数据库的建立与完善，以及各地区、多方面、跨领域的配合，我们能够建立起更加完善、准确的监测预警模型，并且能够把相应的经验应用于除灰飞虱之外的多种虫害的防治预测中。

5 结论

分析结果表明，根据随机森林模型的分析结果我们可以看出，首先，在影响济宁地区灰飞虱发生程度的各气象因子中，“累计60天平均最大风速”与“累计90天平均最大风速”对济宁地区灰飞虱的发生起到了主要作用，所以在中长期有大风且短中期有降水的情况下，有较高的灰飞虱爆发风险。其次，通过训练数据构建的灰飞虱随机森林监测预警模型拥有较高的准确率，随机森林模型袋外数据检验（模型自检）错误率为11.11 % ，利用测试数据进行检验（外部检验），准确率也达88.24 %。最后，由于有关风速的气象因子影响了灰飞虱的迁飞过程，所以造成济宁地区灰飞虱爆发成灾的主要虫源为外地迁飞虫源。除此之外该模型的构建思路不仅仅可以应用于济宁地区灰飞虱的监测预警，对于其他各地区的多种昆虫，都可以采用此类的方法进行相关监测预警模型的构建。综上所述，监测预警模型的科学性和实用性较高，可为植保工作提供参考，也可以在其他地区的多种昆虫的监测预警模型的构建中作为指导；而对于济宁地区的灰飞虱防治工作，除了要预防本地虫源的爆发之外，更要与灰飞虱原迁飞地的防治工作进行配合，以便从根源上减少灰飞虱成灾的风险。

创新点

本研究依照大数据的数据分析思维，较好地实现了大数据在植物保护中的实际应用。实验创新之处主要有以下几点：一是通过本次实验，实现了“大数据 + 现代农业”的理念，运用了较为先进、精准的分析方法对麦田灰飞虱的发生进行了分析预测；二是构建了能够分析济宁地区灰飞虱发生规律的随机森林监测预警模型,即建立了关于灰飞虱非线性的区域化精准监测模型，丰富了相关领域的研究；三是探究了不同时间尺度的各种气象因子对于灰飞虱发生程度的影响，并反映了不同大小的同一气象因子对于灰飞虱发生程度的不同影响力；四是为山东地区小麦、玉米主要虫害特征数据库进行了补充，为灰飞虱的预测预报与科学防治，以及相关植保工作的开展提供了参考和建议。

## 参考文献

包云轩, 田琳, 谢晓金, 陆明红, 姜玉英. 基于大气环流特征量的白背飞虱发生程度短期预报模型. 中国农业气象, 2014, 35(4): 440 - 449.

陈颖. 大数据发展历程综述.当代经济, 2015, 8: 13 - 15.

董师师, 黄哲学. 随机森林理论浅析. 集成技术, 2013, 2(1): 1 - 7.

高东明, 秦文胜, 龚林根, 等. 灰飞虱在条纹叶枯病抗、感品种上的越夏饲养试验. 植物保护, 1994, 20(3): 21.

顾国华, 葛红, 徐莉, 季桦, 薛爱芳. 基于不同气温类型的一代灰飞虱发生期预测. 大麦与谷类科学, 2008, (1): 42 - 45.

胡英华, 王淑霞, 苏加岱. 灰飞虱发生规律及预测预报技术研究. 中国植保导刊, 2011, 31(11): 38 -42.

黄锋.贵州金融城国家级大数据金融中.当代贵州, 2016, 50: 34 - 35.

孔晓民, 蒋飞, 曾苏明, 宋春林, 韩成卫, 吴秋平. 灰飞虱发生消长规律与播期调控玉米粗缩病研究. 作物杂志, 2013(5): 84 - 89.

李济宸, 李桂珍, 高立起, 李青松.灰飞虱发生规律的研究.北京农业科学, 1998.16(4): 24 - 27.

李磊, 牟少敏, 林中琦. 随机森林在棉蚜虫害等级预测中的应用. 安徽农学通报, 2017, 23(1) : 18 - 20.

李欣海. 随机森林模型在分类与回归分析中的应用. 应用昆虫学报, 2013, 50(4): 1190 - 1197.

刘向东, 翟保平, 胡自强. 高温及水稻类型对灰飞虱种群的影响. 应用昆虫学报, 2007, 44(3): 348 - 352.

刘向东, 翟保平, 刘慈明. 灰飞虱种群暴发成灾原因剖析. 应用昆虫学报, 2006, 43(2): 141 - 146.

罗智心, 任荔荔, 祁力言, 周曙东, 戴华国. 温度对桔小实蝇种群发育的影响. 生态学杂志, 2009, 28(5): 921 - 924.

浦茂华. 苏南灰稻蝨（*Delphacodes striatella* Fallén）的初步研究. 昆虫学报, 1963，(2): 3 - 22.

石礼娟,卢军. 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, 2017, 48(1): 169 - 174.

苏加岱, 黄九柏, 刘汉舒, 等. 黄淮海流域玉米粗缩病发生原因分析及防治对策. 山东农业科学, 2009, 9: 59 - 61．

苏加岱, 黄九柏, 姚景勇, 等. 玉米播期与玉米粗缩病发生关系的研究． 现代农业科技，2008, 24: 117 - 120．

孙广仲, 陈志清, 郁祖良, 张大友. 灰飞虱传播的病毒病发生流行特点及耕作与栽培措施的调整对策. 上海农业科技, 2006, 2: 108 - 109．

孙兴全, 吴静菊, 吴爱忠, 支月娥. 灰飞虱生物学特性研究. 上海交通大学学报（农业科学版）, 2000, 18(2) : 150 - 154.

孙增圻, 张再兴.智能控制的理论与技术.控制与决策, 1996, 1: 1 - 8.

唐洪明.大数据背景下企业集团财务管理新趋势.管理观察, 2014, 11: 54 - 55.

汪恩国. 灰飞虱种群数量变动规律与模型测报技术研究. 植物保护, 2007, 33(3): 102 - 107.

王光宏, 蒋平. 数据挖掘综述.同济大学学报(自然科学版), 2004, 32: 246 - 252.

王惠中, 彭安群. 数据挖掘研究现状及发展趋势. 工矿自动化, 2011, 37(2): 29 - 32.

王丽, 韩超, 徐艳博, 蔡广成, 孙友武, 胡学友, 张孝羲, 翟保平. 安徽江淮地区灰飞虱的春季迁飞与扩散. 应用昆虫学报, 2011, 48(5): 1288 - 1297.

王丽爱, 马昌, 周旭东, 訾妍, 朱新开, 郭文善. 基于随机森林回归算法的小麦叶片spad值遥感估算. 农业机械学报, 2015, 46(1): 259 - 265.

王瑞, 翟保平, 孙晓洋. 麦田一代灰飞虱（*Laodelphax striatellus* Fallén）若虫的时空分布. 生态学报, 2007, 27(11): 4536 - 4546.

王彦华, 吴长兴, 赵学平,等. 灰飞虱对杀虫剂抗药性的研究进展. 植物保护, 2010, 36(4): 29 - 35.

温孚江. 农业大数据研究的战略意义与协同机制. 高等农业教育, 2013, 11: 3 - 6.

温孚江. 农业大数据与发展新机遇.中国农村科技, 2013, 10: 14 - 14.

谢润梅.农业大数据的获取与利用. 安徽农业科学, 2015, 30: 383 - 385.

徐艳博, 王孟伦, 朱金生, 韩兰芝，陈法军. 北方水稻和小麦种植区麦田灰飞虱种群发生动态及其带毒率研究. 应用昆虫学报, 2013, 50(1): 215 - 222.

杨波, 刘勇, 牟少敏, 温孚江. 大数据背景下山东省二代玉米螟发生程度预测模型的构建. 计算机研究与发展, 2014, (S2), 160-165.

杨波, 刘勇, 牟少敏,等.大数据背景下山东省二代玉米螟发生程度预测模型的构建.计算机研究与发展, 2014, s2: 160 - 165.

张海燕, 刁永刚, 杨海博, 赵悦, 张孝羲, 翟保平. 山东济宁灰飞虱春季种群动态及迁飞特性. 应用昆虫学报, 2011, 48(5): 1298 - 1308.

张雷, 王琳琳, 张旭东, 刘世荣, 孙鹏森, 王同立. 随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例. 生态学报, 2014, 34(3): 650 - 659.

张晓燕, 翟一凡, 李童, 林清彩, 李丽莉, 张安盛, 门兴元, 吴文虎, 陶玫, 于毅. 不同温度对灰飞虱生长发育的影响. 山东农业科学, 2014,46(12): 99 - 102.

张孝羲. 昆虫迁飞的类型及生理、生态机制. 应用昆虫学报, 1980(5): 46 - 49.

赵北庚. 基于R语言randomforest包的随机森林建模研究. 计算机光盘软件与应用, 2015,(2): 152 - 153.

邹祎.数据挖掘技术综述.信息通信, 2016, 12: 164 - 165.

Breiman L. Random Forests. Machine Learning, 2001, 45(1): 5 - 32.

Breiman L. Bagging Predictors. Machine Learning, 1996, 24(2): 123 -140.

Breiman L. Classification and regression trees. belmont,ca wadsworth international group. Biometrics, 40(3), 17-23.

Chan C W, Paelinckx D. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing of Environment, 2008, 112(6): 2999 - 3011.

Cutler, D. R., Jr, E. T., Beard, K. H., Cutler, A., Hess, K. T., & Gibson, J., *et al*. Random forests for classification in ecology. Ecology, 2007, 88(11): 2783.

Friedl M A, Brodley C E. Decision tree classification of land cover from remotely sensed data. Remote Sensing of Environment, 1997, 61(3): 399 - 409.

Ho, T. K., The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1998, 20(8), 832 - 844.

Kisimoto R. Synoptic weather conditions inducing long-distance immigration of planthoppers, Sogatella furcifera Horvath and Nilaparvata lugens Stal. Ecological Entomology, 1976, 1(2): 95 - 109.

Lee T M, Markowitz E M, Howe P D, *et al*. Predictors of public climate change awareness and risk perception around the world. Nature Climate Change, 2015, 5(11).

M. Pal. Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 2005, 26(1): 217 - 222.

Moreno-Fernández D, Cañellas I, Barbeito I, *et al*. Alternative approaches to assessing the natural regeneration of Scots pine in a Mediterranean forest. Annals of Forest Science, 2015, 72(5): 569 - 583

Otuka A, Matsumura M, Sanadamorimura S, et al. The 2008 overseas mass migration of the small brown planthopper, Laodelphax striatellus, and subsequent outbreak of rice stripe disease in western Japan. Applied Entomology & Zoology, 2010, 45(2): 259 - 266.

Polishchuk P G, Muratov E N, Artemenko A G, *et al*. Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. Journal of Chemical Information & Modeling, 2015, 49(11): 2481.

Qi Y, Kleinseetharaman J, Barjoseph Z. Random forest similarity for protein-protein interaction prediction from multiple sources. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2015, 10: 531 - 542.

Quinlan J R. Induction on decision tree. Machine Learning, 1986, 1(1): 81 - 106.

Quinlan J. R. C4.5: Programs for Machine Learning. Burlington: *Morgan Kaufmanns Publishers*, 1993: 69 - 81.

Quinlan J. R. Induction of decision trees. Machine Learning, 1986, 1(1): 81 - 106.

Seibold S, Brandl R, Buse J, et al. Association of extinction risk of saproxylic beetles with ecological degradation of forests in Europe. Conservation Biology, 2015, 29(2): 382.

Takala-Harrison S, Jacob C G, Arze C, *et al*. Independent emergence of artemisinin Resistance mutations among plasmodium falciparum in Southeast Asia[J]. Journal of Infectious Diseases, 2015, 211(5): 670.

Türkcan S, Masson J B. Bayesian decision tree for the classification of the mode of motion in single-molecule trajectories. Plos One, 2015, 8(12): 1524 - 1528.

Tustison N J, Shrinidhi K L, Wintermark M, et al. Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (Simplified) with ANTsR. Neuroinformatics, 2015, 13(2): 209 - 225.

Wang R, Kwong S, Wang X Z, *et al*. Segment Based Decision Tree Induction With Continuous Valued Attributes. IEEE Transactions on Cybernetics, 2015, 45(7): 1262 - 1275.

Were K, Bui D T, Ystein B. Dick, *et al*. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators, 2015, 52: 394 - 403.

Xia J, Chanussot J, Du P, et al. Spectral–Spatial Classification for Hyperspectral Data Using Rotation Forests With Local Feature Extraction and Markov Random Fields. Geoscience & Remote Sensing IEEE Transactions on, 2015, 53(5): 2532 - 2546.

Zhu Jinliang, Zhu Zengrong, Zhou Yin. Effect of rice sowingdate on occurrence of small brown planthopper and epidemics of planthopper - transmitted rice stripe viral disease. Agricultural Sciencesin China, 2009, 8(3): 332 - 341.

## 致 谢

两年的时光转眼走过，两年的研究生生活即将结束，但是这短短的两年让我学到了很多，懂得了很多。从学习到生活，从理论到实践，从做人到处事各方面都学到了很多，心中充满感激。本研究从定题、课题设计到写作定稿，都是在我的导师刘勇教授的亲切关怀和悉心指导下完成的。他科学态度严肃，治学精神严谨，工作作风精益求精；他学识渊博，对科研事业高度热爱，强烈的求知欲令人折服，是我们学习的指向标。刘勇老师在实验室推行的学术报告会活动使我不仅在学术上有所提高，还锻炼了自己的演讲、汇报以及制作PPT的能力。在此，我要向刘勇老师致以崇高的敬意和诚挚的谢意，感谢他两年以来的指导和教育，正是因为有刘老师一直以来的督促和鼓励，以及对实验方向的把控和对实验内容的严格把关，我的研究才够顺利的走到今天。

衷心感谢我的学校山东农业大学给了我两年深造的机会，提供了良好的学习条件以及学习氛围。感谢温孚江校长将农业大数据引入我校，感谢张宪生校长接力农业大数据，感谢山东农业大学植物保护学院的许永玉教授、薛明教授、刘永杰教授、周成刚教授、郑方强副教授、王宁新副教授、叶保华副教授、刘玉升副教授、谢丽霞老师、张婷婷老师、陈珍珍老师、赵海鹏老师在试验中给与我的支持与帮助。感谢研究生辅导员叶保华老师和教学秘书孟胜国老师给予的工作上的指导。

感谢师兄刘英杰、纪祥龙、梁晓辰、徐润东、徐衡，师姐迟宝杰、杨波、赵雷、张晴晴在学习、工作以及生活中给与的帮助，感谢王康、李洁、李达，师弟师妹战一迪、龚经理、王璐、张宇鑫、任艳平、高珊珊、田泽浩在试验中给予的帮助，谢谢你大家陪伴我度过了人生中意义非凡的两年。

在此，我还要特别感谢我的父母及家人，感谢他们对我的支持和鼓励，为我的研究生生涯提供坚强的后盾。

感谢在百忙之中参加我的论文评审以及答辩工作的各位专家们。

最后，再次向所有给与我帮助的老师、同学、父母、家人以及参加我的论文评审、答辩工作的专家致以诚挚的谢意！

汪 深

2017年6月3日

附 录

表1 本文涉及的所有数据因子

Table 1 All data factors involved in this paper

|  |  |
| --- | --- |
| 输入变量  Input variables | 输出变量  Output variables |
| 平均日降水量、平均气压、平均风速、平均温度、平均水汽压、平均相对湿度、平均20-20时日照时数、平均气压日较差、平均最大风速、平均气温日较差、时间尺度内有无极端温度天气  Average daily precipitation, Average wind speed, Average temperature, Average water vapor pressure, Average relative humidity, Average sunshine hours from 20 to 20, Daily range of Average atmospheric pressure, Average maximum wind speed, Daily range of average temperature, Is there any extreme temperature weather in the time scale | 灰飞虱的发生程度  The occurrence degree of *Laodelphax striatellus* |