

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

杨波 刘勇 牟少敏 温孚江*

(山东农业大学大数据研究中心 山东泰安 271018)

摘要 大数据技术重塑了我们周围的世界。农业大数据是大数据的重要组成部分，是大数据理念、技术和方法在农业领域的应用和发展。依靠多年农作物害虫发生的数据积累，使数据转化为监测预警的行动，是大数据思想和理念的具体体现。通过对山东省 2003-2013 年的平均气温、降水、日照时数、风速等气象资料的相关分析，采用多元线性回归方法，分析预测 6 月下旬、7 月上、中、下旬和 8 月上旬，玉米螟的发生程度，建立了玉米螟动态气候预测模型。经检验，玉米螟的发生与当地的气象条件有直接的关系，6 月下旬、7 月上旬、7 月下旬和 8 月上旬模型拟合的准确率分别为 91%、77%、69%、和 68%。

关键词 农业大数据；玉米螟；监测预警；模型

Based on Big Data: The Establishment of Meteorological Forecast Model for the Occurrence Degree of the Second Generation of Corn Borer in Shandong

Yang Bo Liu Yong Mu Shaomin Wen Fujiang

(Big Data Research Center of Shandong Agricultural University, Taian 271018, Shandong, China)

Abstract Big data technique is reshaping the world around us. Agricultural Big data is an important part of Big data, it is the application and development of Big data's theory, concept and technique in agriculture field. Based on the long-time data accumulation of the occurrence of crop pest, and make these data into practice of monitoring and forecasting, is the embodiment of Big data's theory and concept. The occurrence of corn borer could be closely related to meteorological factors. So the meteorological data including average temperature, precipitation, photoperiod, air speed and so on from 2003 to 2013 in Shandong province were adopted to predict the occurrence of the insect pest depending on multiple linear regression analysis. The dynamic climate prediction models of corn borer in late June, in early, mid and late July, and in early August were established. After testing the model, it was confirmed that occurrence of the corn borer have a direct relationship with the local climate condition, in late June and in early July, in late July and early August, the model accuracy are 91%, 77%, 69%, and 68% respectively.

本课题得到山东省农业重大应用技术创新课题(2014)资助。杨波，女，硕士研究生，Email: xiaobobo2tengzhou@126.com。刘勇，男，教授，Email: liuyong@sdaa.edu.cn。牟少敏，男，教授，Email: msm@sdaa.edu.cn。温孚江*，男，教授，fjwen@sdaa.edu.cn。通讯作者

Key words agricultural big data; corn borer; monitoring and forecasting; model

1 引言

随着互联网、物联网和云计算等网络与通信技术的迅猛发展,由此产生的数据信息和知识比历史上任何时期都要多、都要快,爆炸式增长的数据带领我们迎来了大数据(Big data)时代。大数据的涌现不仅改变着人们的思维方式,也改变着人们的工作和生活方式,甚至改变了企业的运作模式和科技工作的管理和发展方式^[1]。

大数据的数据搜集、分析和挖掘功能,可以发现数据间隐藏的、不容易被人所洞察的相关性,对其进行分析,并通过合适的可视化技术展示出来,这些都有助于人们揭示事物发生和发展的某些现象或规律,做出更快和更符合实际的预测^[2,3]。例如,MIT研究者通过收集和分析在网上销售的或做广告的商品价格,能以很高的准确性估计通货膨胀率。这一方法就可以比传统方法更早地检测通货膨胀的出现^[4]。依大数据的思想理念,荷兰科学家基于多年数据,2014年在《Nature》发文分析指出,新烟碱类杀虫剂吡虫啉的应用,是本地食虫鸟类种群数量减少的主要原因^[5];本年度美国和英国科学家也在《Science》发文,通过分析温度变化与埃塞俄比亚和哥伦比亚高海拔人口密集区的疟疾传播蚊虫和病原的数量变化的关系,指出,全球变暖会导致非洲和南美洲高原地带疟疾病的流行^[6]。

作为大数据重要组成部分的农业大数据,是大数据理念、技术和方法在农业领域的应用和发展,它涉及到农业领域的各个环节,也涉及到生物、气候、地理环境、人口等农业相关领域。采集、整合、利用农业大数据,能够带来现代农业在生产经营、产品流通、市场交易、科研推广和管理决策等方面的变革。农业大数据有其自身特有的复杂性和特殊性:相对于采用二维表来逻辑表达的关系型数据结构,农业领域更多的是半结构化和非结构化数据,如大量的文字、图表、图片、动画语音、视频等形式的组成的超媒体要素,以及专家经验和知识农业模型等^[3]。这些特性都使其更适合应用大数据技术。加之物联网技术向农业各领域渗透,大数据技术在农业上的应用已成为农业信息技术发展的必然趋势^[7]。

植物保护是农业生产的重要组成部分。对农作物病虫害的监测预警是植物保护工作的核心。其重要作用是指导农作物病虫害的科学防控,保障国家粮食和农产品安全。虽然目前在病虫害监测预警方面已有一些专用的技术和软件,但往往采集数据指标偏少,造成预警结果不准确,与实际发生匹配度不高。因此,依多年农作物病虫害发生期作物生长发育状况、气象条件、周边环境、土壤特点、农事管理措施等数据的支撑,以大数据的理念、技术及其预测分析功能,针对病虫害发生趋势和程度进行监测预警,将是未来的发展趋势。

玉米螟 *Pyrausta nubilalis* Hubern, 属鳞翅目(Lepidoptera), 螟蛾科(Pyralidae), 是世界玉米生产的主要害虫之一。在我国已知有亚洲玉米螟[*Ostrinia furnacalis* (Guenée)]和欧洲玉米螟[*Ostrinia nubilalis* (Hübner)]2

种, 其中, 亚洲玉米螟是优势种, 在我国玉米主产区发生 2-3 代, 以第二代为主害代。玉米螟以其幼虫蛀食玉米心叶、茎秆和果穗, 造成玉米作物风折、早枯、缺粒、瘪瘪等现象, 不仅影响玉米的品质, 而且会造成玉米产量的严重损失。本文基于大数据分析的基本理念, 根据已有的数据积累, 选取 9 类气象因子, 以山东省 2003-2013 年 11 年中玉米螟发生程度的实测数据为依托, 构建了以旬为时间尺度的玉米螟发生程度的预测模型。通过相关气象因子对二代玉米螟的发生进行监测预警, 一方面为开展以及优化玉米螟的综合防控措施提供了科学依据, 另一方面也为相关部门及政府决策提供了数据支持, 为我国高产、优质、高效农业服务。

2 资料与方法

2.1 研究区概况

山东省位于中国东部沿海、黄河下游, 北纬 $34^{\circ}22.9'-38^{\circ}24.01'N$ 、东经 $114^{\circ}47.5'-122^{\circ}42.3'E$ 之间, 是玉米生产和消费大省, 玉米面积和总产均居全国前列, 常年种植面积在 270 万 hm^2 左右^[8]。山东省属暖温带季风气候, 降水集中, 雨热同季, 春秋短暂, 冬夏较长。年平均气温 $11^{\circ}C-14^{\circ}C$, 年平均降水量一般在 550-950 mm 之间, 全年降水量 60%-70% 集中于夏季, 光照时数年均 2290-2890 h, 自然条件非常适合玉米生长。玉米螟第二代幼虫盛发期于 7 月中下旬至 8 月上旬, 此时正值夏玉米心叶期与春玉米的乳熟期, 幼虫危害严重影响玉米的品质和产量^[9], 因此, 对二代玉米螟的监测预警成为指导有效防治的关键。

本文选定泰安、临沂、济宁、烟台和滨州代表山东省 5 个不同的玉米种植生态区, 每地市至少选取 5 个采集区(图 1), 第二代玉米螟的发生程度及气象因子数值的确定, 均以这 5 个地区的平均值来代表^[5]。

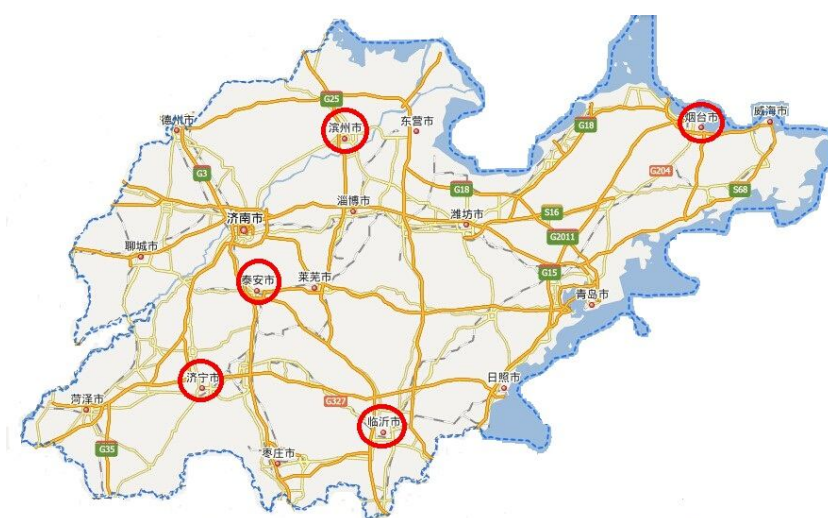


图 1 玉米螟发生数据采集区分布图

2.2 数据来源

山东省 2003-2013 年二代玉米螟发生程度(y)的相关资料来自山东省各植保部门；该时期的逐日气象观测资料来自国家气象信息中心。

2.3 玉米螟发生程度分级

根据中华人民共和国农业行业标准《玉米螟测报技术规范》^[10]，山东省二代玉米螟以百株虫量为指标，发生程度分为 5 级，即轻发生(1 级)、偏轻发生(2 级)、中等发生(3 级)、偏重发生(4 级)、大发生(5 级)。见表 1。

表 1 玉米螟发生程度等级

程度分级	1 级	2 级	3 级	4 级	5 级
百株虫量 Y (头/百株)	$Y \leq 50$	$50 < Y \leq 100$	$100 < Y \leq 200$	$200 < Y \leq 300$	$Y > 300$

2.4 方法

2.4.1 数据处理

将 2003-2013 年 5 地市二代玉米螟的发生资料(word 格式)和影响玉米螟发生的所有气象因子(文本文档格式)按照统一的格式输入到 Excel 表中。选定玉米螟的发生程度和气象资料用来建模^[11]。

根据二代玉米螟的生物学特性，本文采用的影响二代玉米螟种群数量的气象因子是 2003-2013 年的 6 月下旬、7 月上旬、7 月中旬、7 月下旬和 8 月上旬的降水量(x_1)、极大风速(x_2)、平均气压(x_3)、平均风速(x_4)、平均气温(x_5)、平均水汽压(x_6)、平均相对湿度(x_7)、日照时数(x_8)、最大风速(x_9)，分别计算出每年每旬每个气象因子的平均数。

2.4.2 分析方法

首先用二元定距变量的相关分析选出与玉米螟发生程度相关性较好的气候因子且通过显著性检验，再用多元线性回归分析法建立二代玉米螟发生程度的预测模型，最后对模型进行回测检验，测试模型预测效果^[12]。

2.4.3 统计方法

本文所有数据分析均使用 SPSS16.0 统计软件。

3 结果与分析

3.1 以旬为时间尺度的预测模型构建

3.1.1 模型建立

应用二元定距变量的相关分析，可得 9 个气象因子在 6 月下旬、7 月上旬、7 月中旬、7 月下旬和 8 月上旬与二代玉米螟发生程度的相关性大小，影响二代玉米螟发生程度最显著的气象因子在各旬之间是不同

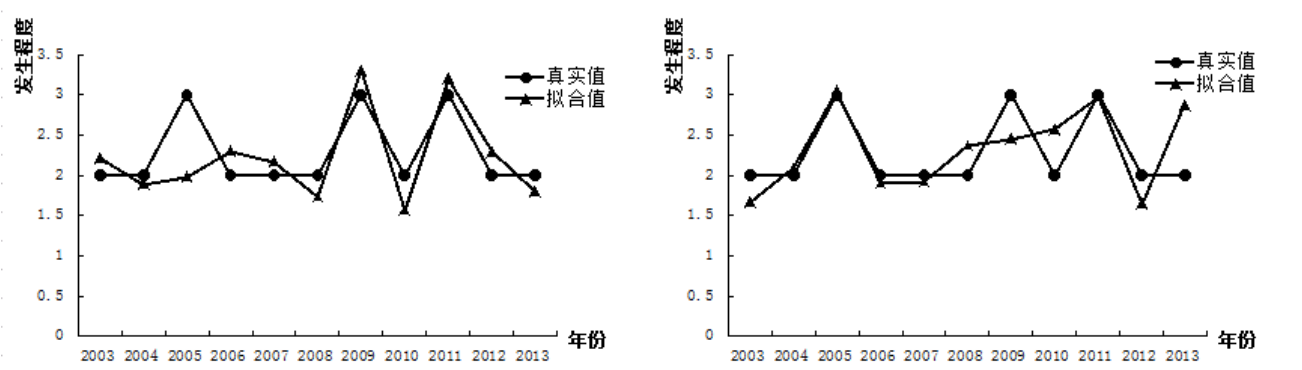
的：6月上旬影响最大的气象因子是平均水汽压和平均相对湿度；7月上旬影响最大的是平均相对湿度和最大风速；7月中旬和7月下旬分别为平均风速、平均相对湿度和降水量和平均风速；平均水汽压和日照时数是8月上旬影响最大的气象因子。以玉米螟发生程度为因变量(y)，以各旬对变量影响程度最大的气象因子为自变量，利用多元线性回归分析，建立各旬的回归分析模型(表2)。

表 2 气象因子对各旬玉米螟发生程度影响的相关性和预测方程

因子名称	6月下旬	7月上旬	7月中旬	7月下旬	8月上旬
降水量(x_1)	0.105	0.022	-0.194	-0.367	0.100
极大风速(x_2)	0.228	-0.201	-0.288	-0.218	-0.197
平均气压(x_3)	0.120	0.199	0.256	0.102	0.140
平均风速(x_4)	0.254	-0.271	-0.434	-0.310	-0.119
平均气温(x_5)	0.103	0.252	0.008	0.189	-0.241
平均水汽压(x_6)	-0.350	-0.195	-0.341	0.098	-0.327
平均相对湿度(x_7)	-0.255	-0.423	-0.441	-0.248	-0.045
日照时数(x_8)	-0.005	-0.100	0.160	0.116	-0.278
最大风速(x_9)	0.082	-0.302	-0.355	-0.280	-0.209
预测方程	$y_1=4.305-0.014x_6$ $y_2=4.807-0.027x_7$ $y_3=5.371-0.025x_4$ $y_4=2.549-421900x_1$ $y_5=4.063-0.007x_6$ $+0.011x_7$ $-0.010x_9$ $-0.03x_7$ $-0.012x_4$ $-0.003x_8$				

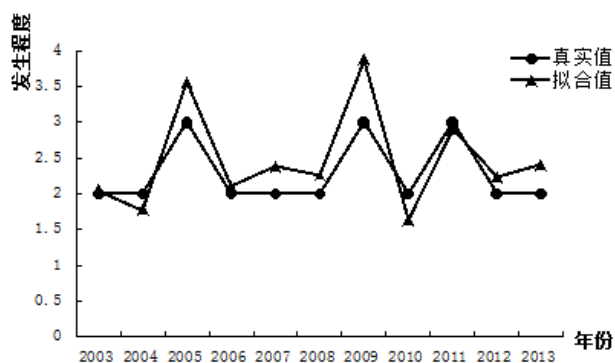
3.1.2 模型检验

利用各旬建立的预测方程对 2003-2013 年山东省二代玉米螟发生程度进行回测检验^[13]，把真实值和拟合值进行对比分析，如图 2。

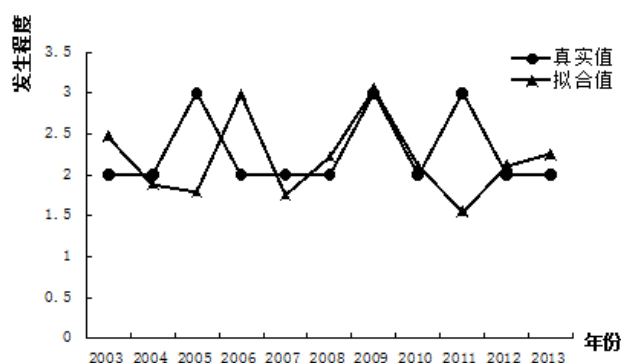


6月下旬玉米螟发生程度真实值与拟合值的对比

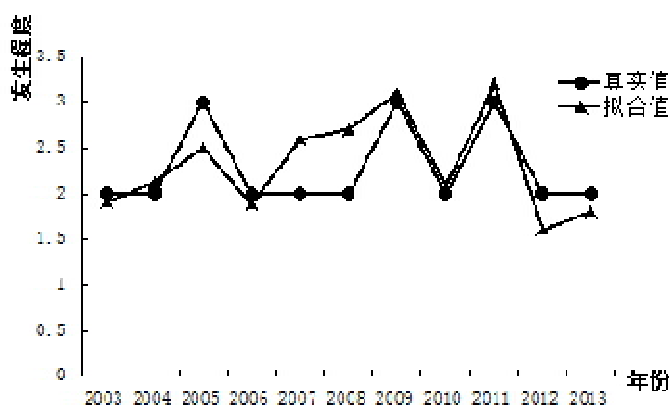
7月上旬玉米螟发生程度真实值与拟合值的对比



7月中旬玉米螟发生程度真实值与拟合值的对比



7月下旬玉米螟发生程度真实值与拟合值的对比



8月上旬玉米螟发生程度真实值与拟合值的对比

图2 山东省二代玉米螟预测方程的回测检验

从图2可以看出,2003-2013年真实值与拟合值逐年变化趋势大致吻合,对于6月下旬的预测,只有2005年拟合值为2,实际值为3,相差1级,准确率91%,表明该模型模拟准确率高;对7月上旬的预测,2003-2007年曲线走势基本吻合,但2008、2009、2010和2013年的预测都相差1级,预测准确率为77%;7月中旬,除2003、2004、2006、2011拟合比较准确,2010年拟合稍微偏低外,其他年份的拟合均值普遍偏高,模型预测准确率仅为36%;7月下旬的预测准确率为69%,2005年和2011年拟合值偏低,2006年偏高;8月上旬预测结果,2005年和2012的拟合值比真实值偏低,而2007年和2008年的拟合值则偏高,准确率68%。

4 结果与讨论

大数据的深入挖掘和利用是新兴产业。对大数据特别是农业大数据的研究,目前还没有建立一套较完备的理论体系;而且针对农业不同的领域和结合点,也缺少高效快速的处理、分析与挖掘的算法与范式。我们尝试把大数据思想与害虫的监测预警结合起来,旨在让大数据理念尽快落地生根,开花结果。

气象因子是影响二代玉米螟种群发生和发展的因子。目前为止,国外对于玉米螟监测预警的研究,主

要集中在越冬存活率、卵孵化率、生长发育水平等方面, Maiorano 在意大利北部, 以每日积温为基础, 构建了玉米螟生理方面的物候模型, 并讨论了玉米螟生长发育与温度的关系^[14]。而国内, 主要是依靠温度和降雨对该虫进行预测研究。袁福香等结合冬季的平均气温、5 月份的平均气温、5 月份的降水量、6 月份的降水量和 7 月上旬的水热系数(降水量 / 平均气温)5 种气象因子建立了吉林省一代玉米螟发生的气象适宜程度等级预报方程, 经实例验证, 中偏重发生年的回代准确率为 71.4%^[15]; 高迎娟等利用吉林省通化市 1981-2007 年玉米螟虫害观测数据和 5 月降水、7 月降水、5 月气温和 5-6 月降水为选定的气象预报因子, 通过对比分析, 并将预报量和预报因子各等级合理赋值, 通过统计手段建立了玉米螟发生程度气象等级预测预报模型, 模型回代检验准确率为 81.5%^[16]。

本文通过对气象因子逐旬进行相关分析可知, 即使在同一个月上、中、下旬, 影响玉米螟发生的主要气象因子是不同的。与以往结果相异的是, 像平均风速、最大风速、平均水汽压等平时不易被纳入考虑范围的气象因子, 竟成了影响玉米螟发生的重要因素, 如 7 月中旬和 7 月下旬玉米螟的发生均与平均风速有关; 而平时经常被提到的平均气温, 在此时间段, 并没有表现出特别大的影响。

在对玉米螟发生程度预报模型检验的基础上,发现模拟结果与实际情况还是有一定差距, 推测原因主要有两点: (1)历史数据的限制。过去十年采集的数据量小、数据范围比较窄。2014-2015 年, 我们扩展了数据采集面, 对农田中诸如生物因子包括自然天敌、周边生物环境等, 非生物因子包括土壤性质等以及地块类型、施肥、浇水等, 都列于采集范围之内。随着数据采集的正规和完善以及数据分析方法的逐步改进, 将会使预测结果与实际情况更加接近。(2)目前建立模型所采用的方法是线性回归法, 这样造成与玉米螟发生程度有关的非线性相关因子就未被考虑, 这在一定程度上影响了预测因子的选取, 也间接影响了预测的准确性^[9]。

大数据的发展日新月异, 利用大数据获得的研究成果也有目共睹, 但大数据分析的确也存在缺陷和偏差问题。因此, 我们在利用大数据分析时, 对于因果关系的判断, 有两点值得注意, 一是有关联不代表有因果关系, 二是一些伪关联情况也可能会造成有因果关系的分析结果假象^[4]。例如, 大数据案例“谷歌流感趋势”近期陷入低谷, 错误率高达 90%以上, 不能预测甲型 H1N1 等重大疫情。它的核心逻辑是: 搜索“流感”的人数与实际患症的人数之间存在相关性, 而事实上, 去医院看流感的人有 80%-90%实际没有得流感, 表面的网络搜索行为与可靠的信息来源还存在较大差距。因此, 就目前而言, 相关关系还不足以替代因果关系, 而只是作为其补充^[2]。尽管数据采集与分析方法现阶段还存在这样或那样的问题, 但大数据的应用前景毋庸置疑, 将会是非常广阔。

参考文献

- [1] Wang Yuan-Zhuo, Jin Xiao-Long, Cheng Xue-Qi. Network Big Data: Present and Future. Chinese Journal of Computer. 2013, 36(6): 1125-1138(in Chinese)

- (王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望. 计算机学报, 2013, 36(6): 1125-1138)
- [2] Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science, 2014, 343, 1203-1205
- [3] Sun Zhong-Fu, Du Ke-Ming, Zheng Fei-Xiang, Yin Shou-Yi. Perspectives of Research and Application of Big Data on Smart Agriculture. Journal of Agricultural Science and Technology, 2013, 15(6): 63-71(in Chinese)
- (孙忠富, 杜克明, 郑飞翔, 尹首一. 大数据在智慧农业中研究与应用展望. 中国农业科技导报, 2013, 15(6): 63-71)
- [4] Ni Guang-Nan. The development and application of Big Data. China Academic Journal Electronic Publishing House, 2013, (9): 6-9 (in Chinese)
- (倪光南. 大数据的发展及应用. 信息技术与标准化. 2013, (9): 6-9)
- [5] Hallmann C A, Foppen R P B, van Turnhout C A M, de Kroon H, Jongejans E. Declines in insectivorous birds are associated with high neonicotinoid concentrations. Nature, 2014, 511: 341-343
- [6] Siraj A S, Santos-Vega M, Bouma M J. Altitudinal Changes in Malaria Incidence in Highlands of Ethiopia and Colombia. Science, 2014, 343: 1154-1158
- [7] Sun Zhong-Fu, Du Ke-Ming, Yin Shou-Yi. Development trend of internet of things and perspective of its application in agriculture. Agricultural Network Information, 2010, 5: 5-8 (in Chinese)
- (孙忠富, 杜克明, 尹首一. 物联网发展趋势与农业应用展望. 农业网络信息, 2010, 5: 5-8)
- [8] Wang Qin-Cheng, Cai Lan-Gao, Li Zong-Xin, Liu Xia. Current Status and Development Strategies of Maize Production in Shandong Province. Journal of Maize Sciences, 2006, 14(5): 159-162 (in Chinese)
- (王庆成, 柴兰高, 李宗新, 刘霞. 山东省玉米的生产现状与发展策略. 玉米科学, 2006, 14(5): 159-162)
- [9] Chen Bin. Studies on the Forecasting of the Second-Generation Corn Borer by the Meteorological Factors. Master Dissertation of Shandong Agricultural University, 2007
- (陈斌. 利用主要气象因子对二代玉米螟预测预报研究. 山东农业大学硕士学位论文, 2007)
- [10] Rules for investigation and forecast technology of the Asian corn borer *Ostrinia furnacalis* (Guenée), NY/T 1611-2008
- (中华人民共和国农业行业标准《玉米螟测报技术规范》, NY/T 1611-2008)
- [11] Li Hong-Jun, Liu Hao-Qiang, Yao Ting-Shan, Hu Jun-Hua, Ran Chun, Lei Hui-De, Huang Liang-Lu, Zhang Ping. Study on prediction mathematical model of occurrence quantity and period for oriental fruit fly (*Bactrocera dorsalis*). Journal of Fruit Science, 2010, 27(2): 275-280 (in Chinese)

(李鸿筠, 刘浩强, 姚廷山, 胡军华, 冉春, 雷慧德, 黄良炉, 张萍. 橘小实蝇发生期及发生量预测模型研究. 果树学报, 2010, 27(2): 275-280)

- [12] Jin Shu-Guang, Yang Mao-Fa, Jin Dao-Chao, Yang Hong, Wen Dong-Mei, Zhu Tao. Establishment and Application of Prediction Model for Outbreak Period and Occurrence Degree of *Cnaphalocrocis medinalis*. Guizhou Agricultural Sciences, 2012, 40(3): 115-118 (in Chinese)

(金曙光, 杨茂发, 金道超, 杨洪, 温冬梅, 朱涛. 稻纵卷叶螟发生期及发生程度预测模型的建立与应用. 贵州农业科学, 2012, 40(3): 115-118)

- [13] Sun Yu-Lian. The Model of Climate Prediction Dynamic of Corn Borer in High Cold Regions in Linxia of Gansu Province. Biological Disaster Science, 2012, 35(3): 337-341(in Chinese)

(孙玉莲. 甘肃临夏高寒地区春玉米螟动态气候预测模型. 生物灾害科学, 2012, 35(3): 337-341)

- [14] Maiorano A. A physiologically based approach for degree-day calculation in pest phenology models: the case of the European Corn Borer (*Ostrinia nubilalis* Hbn.) in Northern Italy. Int J Biometeorol, 2012, 56: 653-659

- [15] Yuan Ju-Xiang, Liu Shi, Guo Wei, Hu Yan-Quan. Prediction of Meteorological Suitability Grades for Occurrence of First Generation of Corn Borer in Jilin Province. Chinese Journal of Agrometeorology, 2008, 29(4): 477-480(in Chinese)

(袁福香, 刘实, 郭维, 胡艳全. 吉林省一代玉米螟发生的气象条件适宜程度等级预报. 中国农业气象, 2008, 29(4): 477-480)

- [16] Gao Ying-Juan, Qi Jin, Xu Li-Ping. The Establishment of Meteorological Forecast Model for the Occurrence Degree of Corn Borer in Tonghua, Jilin. Chinese Agricultural Science Bulletin, 2012, 28(21): 211-215(in Chinese)

(高迎娟, 齐金, 徐丽萍. 吉林通化地区玉米螟发生程度气象等级预测预报模型的建立. 中国农学通报, 2012, 28(21): 211-215)