应用昆虫学报 *Chinese Journal of Applied Entomology* 2013,50(4):1170 - 1177. DOI:10.7679/j.issn.2095 - 1353.2013.161

基于广义可加模型的昆虫种群动态非线性分析及 R 语言实现*

欧阳芳 戈 峰**

(中国科学院动物研究所 农业虫害鼠害综合治理研究国家重点实验室 北京 100101)

摘 要 昆虫种群受到气候、天敌和土壤等多种生态因子的综合作用,其动态具有复杂性、不确定性和非线性等特征。广义加性模型(generalized additive models,GAM)就是适用于响应变量与解释变量之间的关系是非线性或非单调的数据分析。本文以 1973—1990 年稻纵卷叶螟种群数量与降雨持续天数和降雨量的相关性分析为例,介绍了广义可加模型的应用及其 R 语言实现步骤,为研究昆虫种群动态及其驱动因子提供了有效的分析工具。

关键词 种群动态,非线性分析,广义可加模型,R 语言

Nonlinear analysis of insect population dynamics based on generalized additive models and statistical computing using R

OUYANG Fang GE Feng**

(State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China)

Abstract Population dynamics are effect by ecological factors such as climate, natural enemies and soil conditions. Nonlinearity, complexity and uncertainty are inherent in insect population dynamics. Generalized additive models (GAMs) are often used to analyse nonlinear or non-monotonic relationships between response and explanatory variables. These models provide effective tools for exploring the dynamics of insect populations and their primary driving factors. This paper introduces the use of generalized additive models in insect population dynamics and their analysis using R. The relationship between the population sizes of rice leaf folder, *Cnaphalocrocis medinalis* Guenee and the duration and amount of precipitation from 1973 to 1990 are modeled using GAM.

Key words population dynamics, analysis, generalized additive models, R project

种群动态是昆虫种群生态学研究的核心问题。自然界中昆虫种群受到气候、天敌和土壤等多种生态因子的综合作用,其动态具有复杂性、不确定性和非线性等特征(赵中华和沈佐锐,2001)。尤其是当前全球变化背景下,分析昆虫种群动态与气候变化、农田格局变化的关系,是进行昆虫预警和害虫可持续治理的基础和前提(戈峰,2011;欧阳芳和戈峰,2011)。广义加性模型为研究昆虫种群数量时空动态规律及其驱动因子提供了有效的工具。本文主要介绍广义加性模型这种非线性分析方法的应用及其 R 语言实现步骤。

1 广义可加模型

Hastie 和 Tibshirani(1986,1990)提出并发展了广义加性模型(generalized additive models, GAM)。此模型适用于响应变量与解释变量之间的关系是非线性和非单调的数据分析(贾彬等,2005)。众所周知,在数理统计分析中,研究变量之间相关关系常用方法有回归分析(regression)。最简单的情况是一元线性回归模型(simple linear regression)。假定有一个响应变量(response variable)Y 和一个解释变量(explanatory variable)

* 资助项目:国家自然科学基金(31200321)和国家科技支撑计划项目(2012BAD19B05)。

**通讯作者,E-mail:gef@ioz.ac.cn

收稿日期:2013-06-08,接受日期:2013-06-30

X,则 Y 的条件期望与 X 的关系可用线性函数 $E(Y/X) = a_0 + a_1X + \varepsilon$ 表示。将这一函数推广到多个解释变量,即多重线性回归模型(multiple linear regression):$E(Y/X_1, \cdots, X_n) = a_0 + a_1X_1 \cdots + a_nX_n + \varepsilon$。然而在很多情况下,$Y$ 的条件期望与 X 并非简单的线性关系,用线性模型拟合数据就不一定适合。此时,可根据具体情况改变两者间的函数形式构建模型来描述它们之间的关系。一种方式是改变响应变量条件期望的函数形式,将它记为 $g(muY)$,其中 $muY = E(Y/X_1, \cdots, X_n)$。模型可表示为:$g(muY) = a_0 + a_1X_1 \cdots + a_nX_n + \varepsilon$,称之为广义线性模型。另一种方式是用非参数的形式来描述响应变量条件期望与解释变量的对应关系,用 $f(x)$ 表示。模型可表示为:$E(Y/X) = f(x)$。将其推广到多个解释变量时,则模型变成 $E(Y/X_1, \cdots, X_p) = a + f(X_1) \cdots + f(X_n) + \varepsilon$,称之为可加模型(additive models)。将两种方式结合起来,则模型可表达为:

$$g(\mathrm{mu}Y) = a_0 + f_1(X_1) \cdots + f_n(X_n) + \varepsilon,$$

即为广义可加模型。

其中,在 $g(\mathrm{mu}Y)$ 项中,$\mathrm{mu}Y$ 是 Y 的期望值,$g(\cdots)$是连接函数,a_0 是截矩,在 $f_n(X_n)$ 项中,$f_n(\cdots)$是解释变量 X_n 的单变量函数,ε 为随机变量。

基于广义可加模型,分析昆虫种群数量时空动态(响应变量)与生态因子(解释变量)之间的相关性,主要目的是明确:1)显著影响种群数量时空动态的关键生态因子;2)种群数量时空动态与关键生态因子是正相关,还是负相关;3)线性是线性相关,还是非线性相关。

2　广义可加模型的分析步骤

广义可加模型拟合响应变量与解释变量之间的非线性关系的过程中,需要同时考虑曲线的拟合优度和拟合光滑度。用模型的确定系数来度量拟合优度,即满足响应变量的预测值和观测值之间残差平方和最小。用响应变量与解释变量之间的关系曲线粗糙度来度量拟合光滑度,即要求得到的拟合曲线平稳变化,而非剧烈快速波动。模型分析可大致分为以下几个步骤:(1)变量预分析,(2)模型构建,(3)模型估算,(4)模型优化。

(1)变量预分析

数据分析之始,首先根据研究目的,区分数据集中的响应变量和解释变量。在昆虫种群动态分析中,种群数量或者种群数量变化率作为响应变量,各种生态因子作为解释变量。其次,利用频次分布图或正态 QQ 图等方法,掌握响应变量的分布特征或分布型以大致确定其连接函数(表 1)。再次,分析解释变量之间的相互关系。类似于线性分析中解释变量之间可能存在的共线性关系,在广义可加模型的非线性分析中,需要克服解释变量之间的共曲线性关系导致的系数标准误的偏差。可以利用两个解释变量之间的 Pearson 相关系数 R 来判别两者的相关性程度。设定当其 $R > 0.5$ 时两个解释变量之间存在共曲线性关系,即在模型构建时,只选取其中一个变量作为解释变量。

表 1　响应变量分布型和连接函数

Table 1　Distribution of response variable and link functions

响应变量分布族 Family of response variable	连接函数 Link functions
正态分布 Normal distribution	Identity link: $g(z) = z$
二项分布 Binomial distribution	Logit link: $g(z) = \log(z/(1-z))$
负二项分布 Negative binomial distribution	Inverse link: $g(z) = 1/z$
伽码分布 Gamma distributions	Log link: $g(z) = \log(z)$
泊松分布 Poisson distributions	Log link: $g(z) = \log(z)$

(2)模型构建

变量预分析基础上,明确响应变量与单个待选解释变量之间的关系。大致判断每个待选解释变量作为模型的参数还是非参数形式,确定分析模型的基本组成,建立初步的模型结构。构建模型时,需要考虑等效模型(equivalent models)。对于一组多变量数据集,构建模型时可构造出一个或多个可行或可能的模型。如以不同解释变量的组合可构成多个分析模型,或者以相同的解释变量和不同的组合形式构建数个不同形式的模型。不同形式的模型是指不同模型的各解释变量的函数形式不同,即单变量函数形式不同。因此初步

构建的响应变量与解释变量之间的分析模型是包括多个可能模型的模型组。如:数据集中有 i 个解释变量,则以不同解释变量组合构建的模型组包括可能模型的数量为: $C_i^1 + C_i^2 + \cdots + C_i^{i-1} + C_i^i$。

模型 1 $g(muY) = a_0 + f_1(X_1) + \varepsilon$

模型 2 $g(muY) = a_0 + f_1(X_1) + f_2(X_2) + \varepsilon$

…

模型 i $g(muY) = a_0 + f_1(X_1) \cdots + f_i(X_i) + \varepsilon$

(3)模型估算

同时考虑曲线的拟合优度和拟合光滑度的前提下,对初步构建的多个模型分别进行估算。估算主要包括连接函数 $g(\cdots)$ 的估计、解释变量 X_n 的单变量函数 $f_n(\cdots)$ 的估计、光滑参数的估计等方面。连接函数的估计有局部积分法(local-scoring procedure)、平均导数法(average derivatives)(Hardle and Stoker,1989)和限制回归法(slicing regression)(Duan and Li,1991)等。单变量函数的估计有移动均数、移动中位数、移动线性、局部加权移动线性光滑函数、核光滑函数(Rosenbla,1971;Hardle and Gasser,1984),和样条函数(Schoenbe and Greville,1965;Reinsch,1967)等。常用的光滑参数估计方法有广义交叉验证的偏差(generalized cross-deviance,GCV)和赤池信息准则(akaike information criterion, AIC)。

(4)模型优化

对模型组估算后,需要筛选出满足预定要求和达到预设目标的优化模型。基于广义可加模型的昆虫种群动态非线性分析,最优化模型的原则设定为:1)回归模型中所有生态因子(解释变量)的影响均达到显著水平;2)在满足所有生态因子的影响均达到显著水平的前提下,以模型评价指标如 GCV 或 AIC 来筛选出最优模型。目前常用的方法是认为 GCV 或 AIC 值最小的模型为最优模型。

3 广义可加模型的 R 语言实现

田间昆虫种群受多种生态因子的综合作用。需要明确哪些生态因子对该种群动态起到关键作用呢? 这种作用是正作用还是负作用,是非线性还是线性的。以表 2 数据为例,现应用广义线性模型,分析 1973—1990 年某市历年稻纵卷叶螟全代累计蛾量与降雨持续天数和降雨量的相关性。

广义可加模型常用的分析工具有 R 软件平台的 mgcv 软件包(Wood,2006)。本文以表 2 数据为例,简单介绍该软件包分析昆虫种群动态与生态因子关系的基本操作。

表 2 某市历年稻纵卷叶螟虫情和部分气象资料

Table 2 Population amount of rice leaf folder, duration and capacity of precipitation in a study region

年份 Year	全代累计蛾量 Adult	降雨持续天数(d) Day	降雨量(mm) Precipitation
1973	27 285	15	387.3
1974	239	14	126.3
1975	6 164	11	165.9
1976	2 535	24	184.9
1977	4 875	30	166.9
1978	9 564	24	146.0
1979	263	3	24.0
1980	3 600	21	23.0
1981	21 225	13	167.0
1982	915	12	67.0
1983	225	17	307.0
1984	240	40	295.0
1985	5 055	25	266.0
1986	4 095	15	115.0
1987	1 875	21	140.0
1988	12 810	32	369.0
1989	5 850	21	167.0
1990	4 260	39	270.8

(1)变量预分析

数据准备: 将表 2 数据保存到文本文件中,命名为:Rice_insect.txt。

R 语言的运行命令如下或见图 1;#后面的文字表示对 R 语言命令的注释。

```
library(mgcv)   # 加载 R 软件中的 mgcv 软件包
Data <- read.delim("Rice_insect.txt")   # 读取原始数据
```

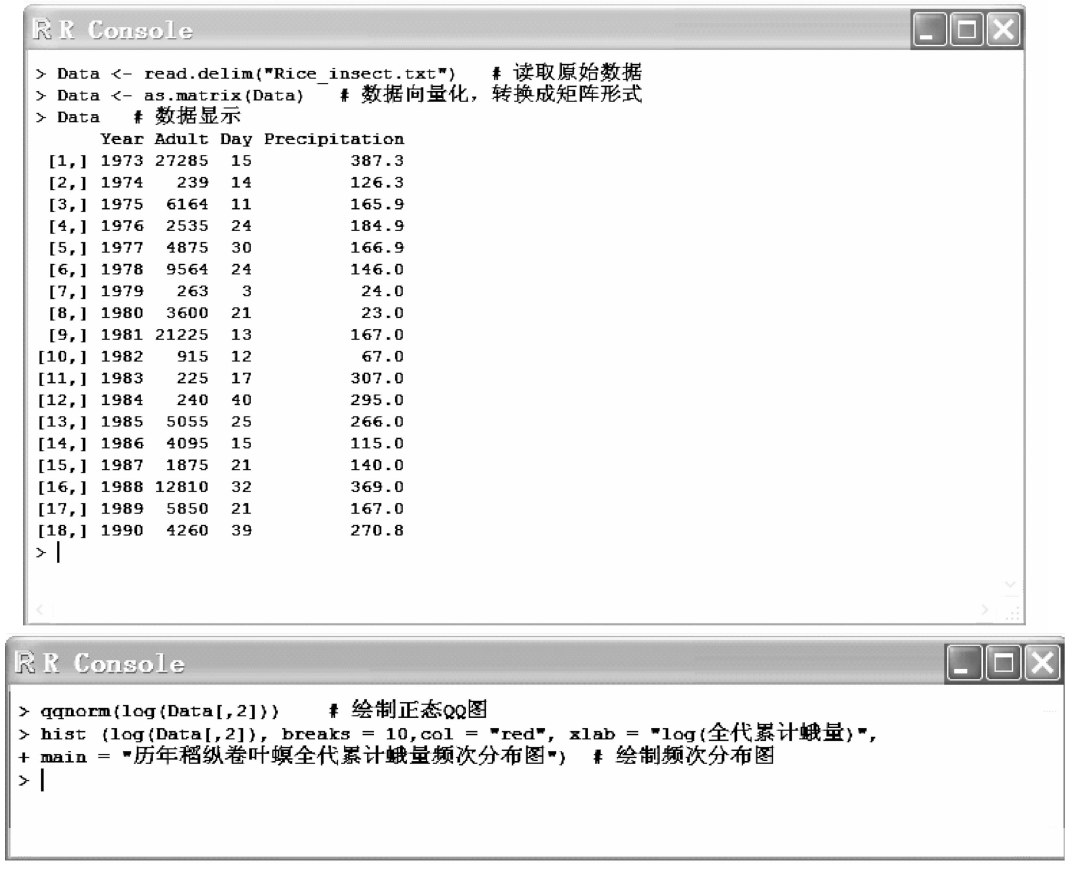

图 1　R 语言的运行命令和显示数据

Fig. 1　Run command in R project and display data

Data < - as. matrix(Data)　# 数据向量化,转换成矩阵形式

Data　# 数据显示

绘制正态 QQ 图和频次分布图(图 2):

R 语言的运行命令如下:#后面的文字表示对 R 语言命令的注释。

qqnorm(log(Data[,2]))　　# 绘制正态 QQ 图

hist (log(Data[,2]), breaks = 10, col = "red", xlab = "log(全代累计蛾量)",

main = "历年稻纵卷叶螟全代累计蛾量频次分布图")　# 绘制频次分布图

根据图 2 正态 QQ 图和频次分布图,初步设定历年稻纵卷叶螟全代累计蛾量(响应变量)的分布特征为正态分布类型,Identity link 作为连接函数。

分析解释变量的相关性:

R 语言的运行命令如下:#后面的文字表示对 R 语言命令的注释。

cor. test(Data[,3], Data[,4], method = "pearson") # 计算两生态因子的相关性,并进行 t 检验。

图 3 结果表明 pearson 相关系数为 0.5127947。根据预设的要求,当 $R > 0.5$ 时候,说明两个生态因子降雨持续天数(d)和降雨量(mm)之间存在共线性相关,因此构建模型的时候,只选择其中一个变量作为解释变量。

(2)模型构建

据变量预分析的结果,构建模型时只选择两个生态因子之一作为解释变量。将稻纵卷叶螟全代累计蛾量 Adult 作为响应变量,降雨持续天数 Day 和降雨量 Precipitation 分别作为解释变量构建模型,模型组:

模型 1　$g(\text{Adult}) = \alpha + f_1(\text{Day}) + \varepsilon$

模型 2　$g(\text{Adult}) = \alpha + f_2(\text{Precipitation}) + \varepsilon$。

(3)模型估算

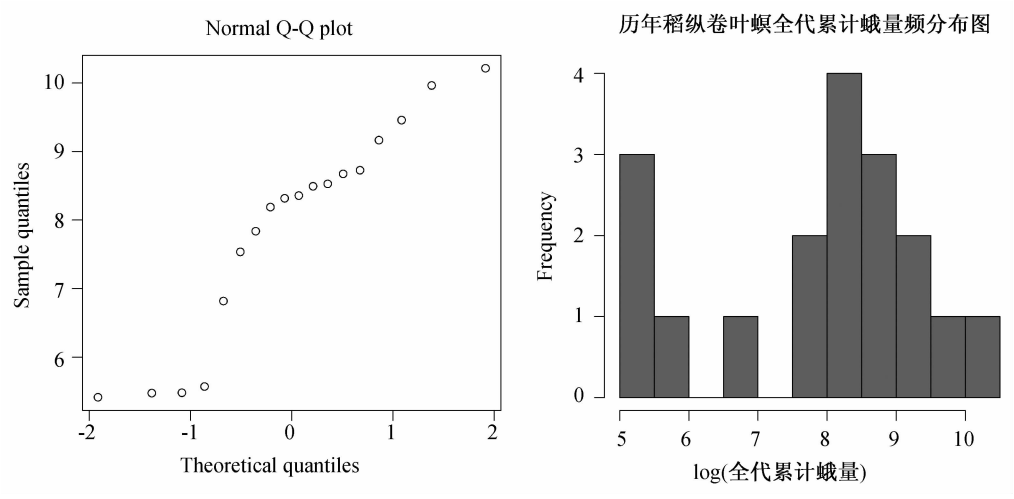

图 2 R 语言的运行命令,正态 QQ 图和频次分布图

Fig. 2 Run command in R project, normal Q-Q plot and frequency distribution diagram

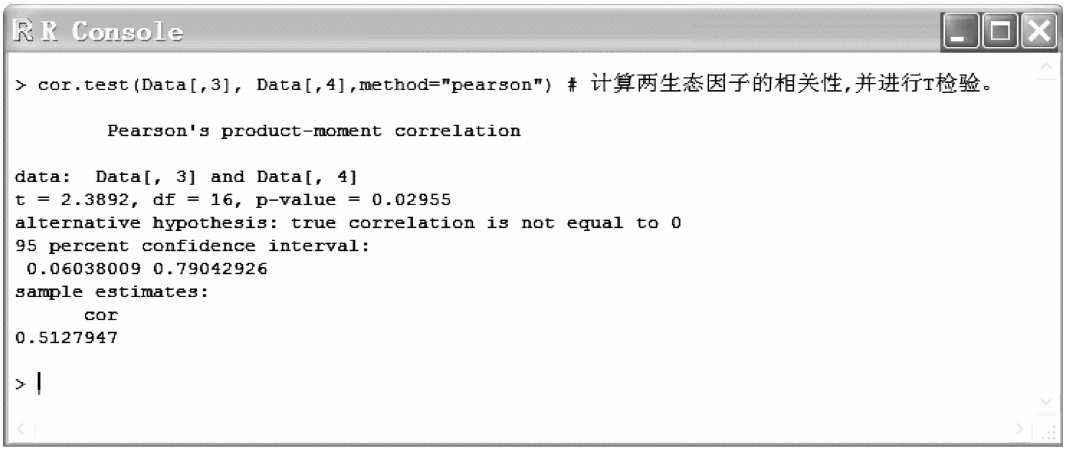

图 3 R 语言的运行命令和 Pearson 相关系数

Fig. 3 Run command in R project and Pearson correlation coefficient

定义变量,将原始数据输入到指定的变量名。将累积量数据输入到变量 Adult,将降雨持续天数输入到变量 Day,将降雨量输入到变量 Precipitation。

在 R 软件 mgcv 软件包下,对模型 1 和模型 2 进行估算。

R 语言的运行命令如下:#后面的文字表示对 R 语言命令的注释。

sult1 < - gam(log(Adult) ~ s(Day)) # 对模型 1 进行估算

summary(Result1) # 输出模型 1 结果

图 4 结果中,生态因子降雨持续天数(Day)的作用的影响水平 $P = 0.489$,说明该因子影响不显著。

sult2 < - gam(log(Adult) ~ s(Precipitation)) # 对模型 2 进行估算

summary(Result2) # 输出模型 2 结果

5 结果中,生态因子降雨量(Precipitation)的作用的影响水平 $P = 0.0772$。说明该因子在 $P < 0.1$ 水平下影响显著。为了介绍方法需要,本文设定 $P < 0.1$ 为显著水平。一般情况设定 $P < 0.05$ 为显著水平。

(4)模型优化

根据对两个模型的估算,模型 1 中降雨持续天数 Day(解释变量)的影响未达到显著水平。模型 2 模型中降雨量 Precipitation(解释变量)的影响

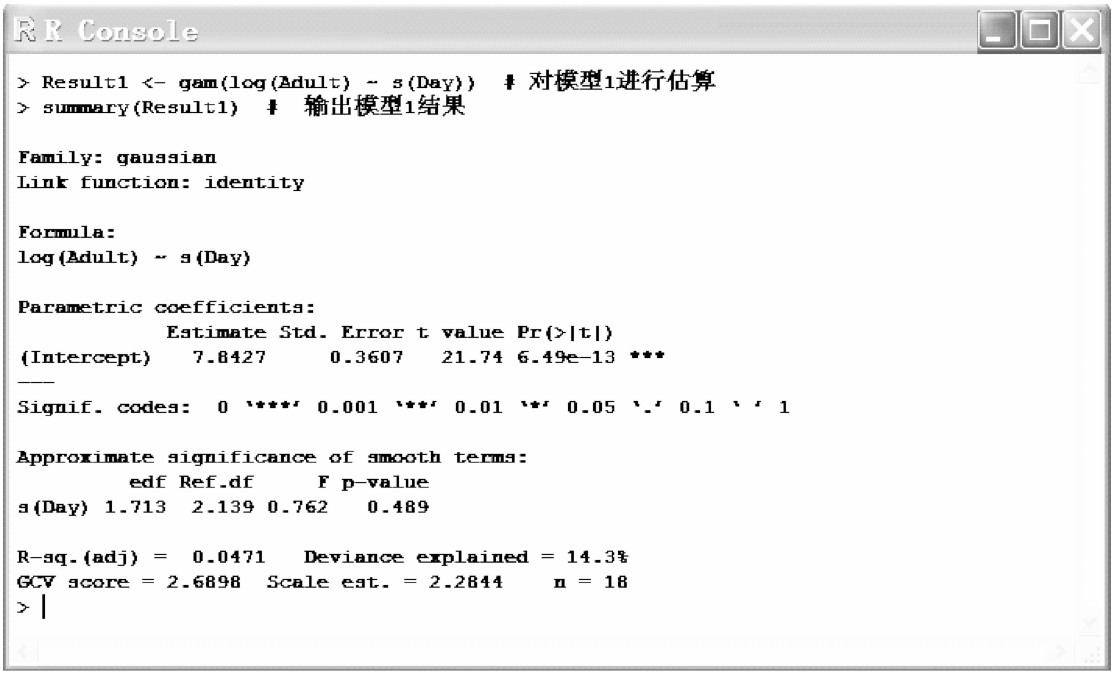

图 4　R 语言的运行命令和模型 1 结果

Fig. 4　Run command in R project and result of modle 1

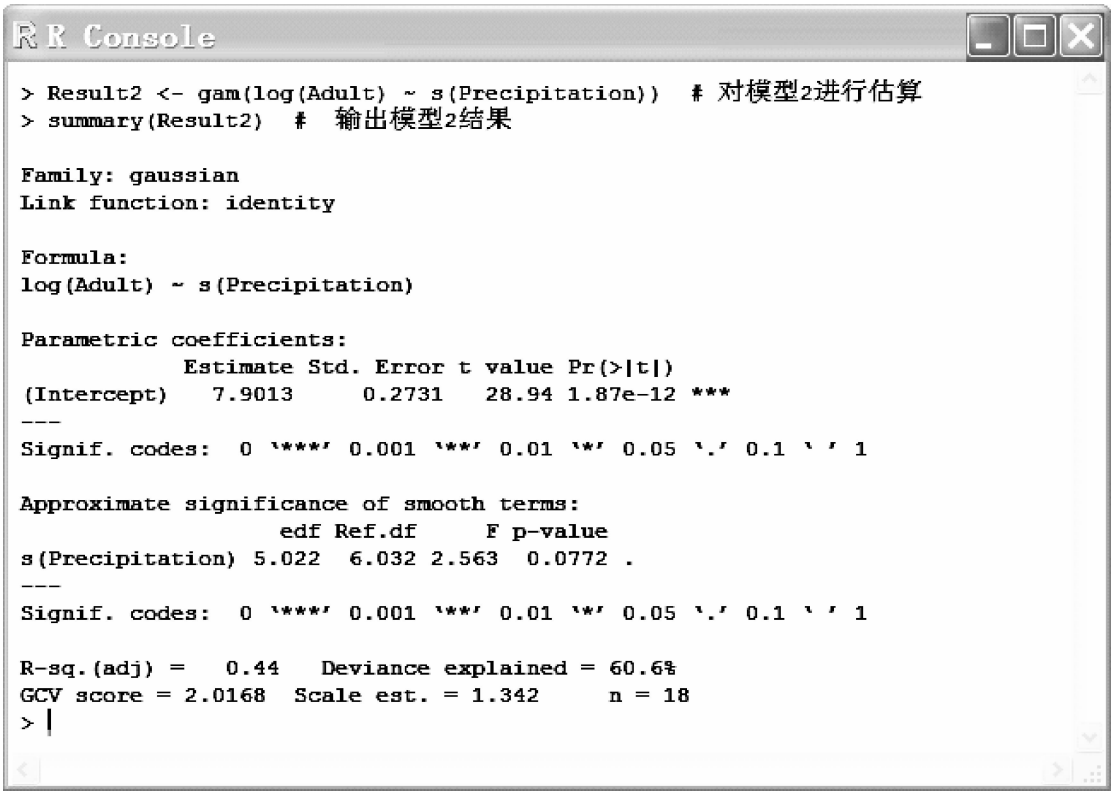

图 5　R 语言的运行命令和模型 2 结果

Fig. 5　Run command in R project and result of modle 2

未达到显著水平。同时比较两个模型的 GCV 值，模型 2 的 GCV score 2.0168 值小于模型 1GCV score 值 2.6898。

因此，确定模型 2 为最优模型。模型 2 分析表明，降雨量 Precipitation 在 $P < 0.1$ 水平下显著影响历年稻纵卷叶螟种群动态。历年稻纵卷叶螟全代累计蛾量与降雨量的相关性如图 6。

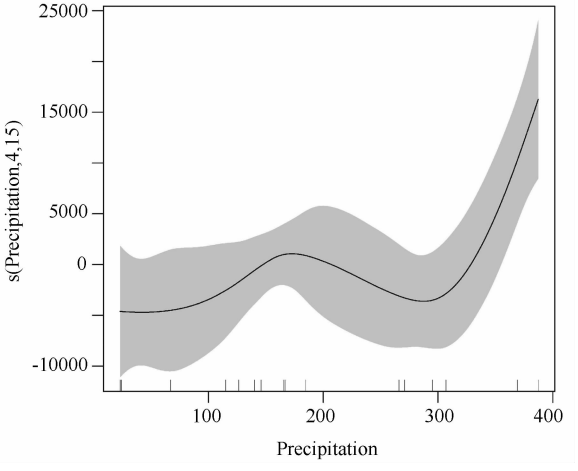

图 6 历年稻纵卷叶螟全代累计蛾量与降雨量的非线性关系

Fig. 6 Nonlinear relationship between population amount of rice leaf folder, *Cnaphalocrocis medinalis* and capacity of precipitation

图 6 表明：历年稻纵卷叶螟全代累计蛾量与降雨量的关系是非线性的，总体来看，是正相关；分阶段来看，将降水量在 180 mm 以下为局部正相关，180～300 mm 为局部负相关，300～400 mm 的为局部正相关。

基于广义可加模型的昆虫种群动态非线性分析，可以得出历年稻纵卷叶螟全代累计蛾量与降雨持续天数和降雨量的相关性结果：1）显著影响种群数量时空动态的关键生态因子有降水量；2）种群动态与降水量之间是非线性相关；3）两者之间的关系总体呈正相关。

4 讨论

本文只介绍了最基本的广义可加模型（GAM）在 R 软件中的使用。广义可加模型（GAM）应用广泛，并受到越来越多的关注。

与广义线性模型（GLM）不同，广义可加模型（GAM）的基本假设是各函数项具有可加性和光滑性，适用于响应变量的期望与解释变量间的关系是非线性和非单调的数据。广义线性模型（GLM）是由模型驱动的，预先假定各解释变量项的参数形式，此形式局限于已知曲线的形状。而广义可加模型（GAM）是由数据驱动的，各解释变量项是由反应变量期望的函数与预测变量关系的曲线形状决定的非参数形式。即对于广义可加模型来说，数据所决定的反应变量与解释变量之间的关系，不是关系的参数形式，而是关系的本质联系（Yee and Mackenzie，1991，2002）。因此，广义可加模型（GAM）具有较高的灵活性，其对数据资料的要求较少，它的适用范围非常广泛。

虽然广义可加模型（GAM）得到了广泛的关注和应用，但其需要克服一定的局限性，如：1）对异常点或者离群点数据敏感；2）其没有考虑解释变量间的交互作用；3）存在共曲线性等。此外，广义可加模型分析体系的数学原理及其运算过程较为复杂。本文仅涉及到初步内容，如需进一步了解和拓展 R 软件平台的应用，可以参考 mgcv 软件包（Wood，2006）。

参考文献（References）

Duan N, Li KC, 1991. Slicing Regression - a link-free regression method. *Ann. Stat.*, 19(2):505－530.

Hastie T, Tibshirani R, 1986. Generalized additive models (with discussion). *Stat. Sci.*, 1(3):297－318.

Hardle W, Gasser T, 1984. Robust non-parametric function fitting. *J. Roy. Stat. Soc. Series B-Methodol.*, 46(1):42－51.

Hardle W, Stoker TM, 1989. Investigating smooth multiple-regression by the method of average derivatives. *J. Am. Stat. Assoc.*, 84(408):986－995.

Hastie T, Tibshirani R, 1990. Generalized Additive Models. Chapman and Hall. 1－335.

Reinsch CH, 1967. Smoothing by Spline Functions. *Numer. Math.*, 10:177－183.

Rosenbla M, 1971. Curve estimates. *Ann. Math. Stat.*, 42(6):1815－1842.

Schoenbe IJ, Greville TN, 1965. Smoothing by generalized spline functions. *Siam Rev.*, 7:617－620.

Wood SN, 2006. Generalized additive models: an introduction with R, CRC. 1－391.

Yee TW, Mackenzie M, 2002. Vector generalized additive

models in plant ecology. *Ecol. Modell.*, 157(2/3): 141 – 156.

Yee TW, Mitchell ND, 1991. Generalized additive-models in plant ecology. *Journal of Vegetation Science*, 2(5): 587 – 602.

戈峰, 2011. 应对全球气候变化的昆虫学研究. 应用昆虫学报, 48(5): 1117 – 1122.

贾彬, 王彤, 王琳娜, 陈长生, 2005. 广义可加模型共曲线性及其在空气污染问题研究中的应用. 第四军医大学学报, 26(3): 280 – 283.

欧阳芳, 戈峰. 2011. 农田景观格局变化对昆虫的生态学效应. 应用昆虫学报, 48(5): 1177 – 1183.

赵中华, 沈佐锐. 2001. 昆虫种群动态非线性建模理论与应用. 生物数学学报, 16(4): 439 – 444.