·专题报道·

ARIMA 模型在黄瓜霜霉病疾病指数时间序列建模中的应用研究

华来庆[1]，申广荣[2]，熊林平[1*]，孟　虹[1]，赵胜荣[3]，胡亚萍[4]
(1.第二军医大学卫生勤务学系卫生统计学教研室，上海 200433；2.上海交通大学农业与生物学院，上海 201101；3.上海市松江区蔬菜技术推广站，上海 201613；4.上海市浦东新区农业技术推广中心，上海 201201)

[摘要] **目的**：探索黄瓜霜霉病疾病指数时间序列预测方法。**方法**：采用黄瓜霜霉病病情指数时间序列进行研究，通过模型识别、残差方差比较、参数估计及其检验、观察参数之间相关系数矩阵、白噪声检验、模型的拟合度分析等过程。**结果**：建立了 *ARIMA*(2,2,0)模型：$(1+0.487\ 1B+0.554\ 7B^2)(1-B)^2y_t=a_t$。*ARIMA*(2,2,0)模型的预测值误差平方和 *SSE*=0.001 822，根均方误差 *RMSE*=0.008 537，且验证数据的预测值与原始值吻合较好。*ARIMA*(2,2,0)模型为本研究获得的预测效果较好的一维时间序列模型，适合于黄瓜霜霉病的中期、后期预测。**结论**：通过残差方差定阶法缩小模型选择范围，再结合模型的参数估计、相关系数矩阵、白噪声检验和拟合优度检验最后确定模型的思路，有利于快速准确找到合适的模型。

[关键词] *ARIMA* 模型；黄瓜霜霉病；疾病指数时间序列
[中图分类号] R 195.1　　**[文献标识码]** A　　**[文章编号]** 0258-879X(2006)07-0729-04

Application of autoregressive integrated moving average model in establishing disease index time series model of cucumber downy mildew disease

HUA Lai-qing[1], SHEN Guang-rong[2], XIONG Lin-ping[1*], MENG Hong[1], ZHAO Sheng-rong[3], HU Ya-ping[4] (1. Department of Health Statistics, Faculty of Health Services, Second Military Medical University, Shanghai 200433, China; 2. School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai 201101; 3. Agro-technical Promotion Station, Songjiang District of Shanghai, Shanghai 201613; 4. Agro-technical Promotion Station, Pudong New District of Shanghai, Shanghai 201201)

[ABSTRACT] **Objective**: To explore the forecasting method of disease index time series of cucumber downy mildew disease. **Methods**: Using the time series of cucumber downy mildew disease, we established an autoregressive integrated moving average model, *ARIMA*(2,2,0) based on model identification, comparison of residual variance, estimation and verification of parameter, observation of the correlation of the estimates matrix, autocorrelation check of the residuals, analysis of the fitting of model and so on. **Results**: An *ARIMA* model (2,2,0) was established: $(1+0.487\ 1B+0.554\ 7B^2)(1-B)^2y_t=a_t$, with the Sum of Squared Error (*SSE*) being 0.001 822 and the Root of Mean Squared Error (*RMSE*) being 0.008 537. The predicted values of validating date fitted well with the primary values. The established model showed satisfactory forecasting ability and was suitable for forecasting the middle stage and late stage cucumber downy mildew disease. **Conclusion**: Limiting the alternatives of model by residual variance, together with parameters estimation, the correlation of the estimates matrix, the autocorrelation check of the residuals and the fitting test, can help to search for suitable model quickly and accurately.

[KEY WORDS] *ARIMA* model; cucumber downy mildew disease; disease index time series
[Acad J Sec Mil Med Univ, 2006, 27(7): 729-732]

黄瓜霜霉病是黄瓜栽培中常见的严重气生真菌性病害，气流传播，侵染频繁，病害发展迅速，每年都给黄瓜生产造成不同程度的损失。掌握霜霉病的发病规律，做好预测预报工作有着极为重要的意义。本研究应用时间序列方法，建立黄瓜霜霉病病情指数 *ARIMA*(2,2,0)模型，探索黄瓜霜霉病疾病指数时间序列建模方法[1]。

1　材料和方法

1.1　*材料*　采用上海市松江区蔬菜基地 2004 年3～6月黄瓜霜霉病疾病指数数据。剔除前期发病株率持续为 0 的数据，共得到 30 个时间点的数据(每 3 d 记录 1 次)，用前 29 个数据建模。

1.2　*方法*　通过对黄瓜霜霉病疾病指数时间序列的平稳性变换并检验、模型识别、模型估计和诊断检查、预测、模型验证过程，建立了 *ARIMA*(2,2,0)模型。使用 SAS/ETS 软件建模。

[基金项目] 上海市科委科技攻关计划(03DZ19314). Supported by Grants for Tackling Key Program of Shanghai Science and Technology Committee(03DZ19314).
[作者简介] 华来庆，硕士.
* Corresponding author. E-mail: xionglinping@yahoo.com.cn

2　结　果

2.1　时间序列的平稳性检验　建立 *ARIMA* 模型的基础是时间序列必须是平稳性的，本研究序列的折线图（略）具有明显的上升趋势，30 个时间点的数据分别为 0.005 0、0.010 0、0.012 5、0.015 0、0.017 5、0.027 5、0.032 5、0.037 5、0.042 5、0.055 0、0.070 0、0.077 5、0.100 0、0.115 0、0.130 0、0.160 0、0.187 5、0.190 0、0.207 5、0.222 5、0.237 5、0.252 5、0.265 0、0.267 5、0.275 0、0.310 0、0.335 0、0.355 0、0.390 0、0.427 5，该序列是一个不平稳的序列。因此考虑对原始序列进行变换，使其达到平稳化。用图示法、*Daniel* 检验法、*Kendall* τ 检验法、自相关函数（*ACF*）检验法、修正的 *Box-Pierce Q* 检验法等多种方法结合起来综合判断，发现两次差分变换相对较为合理。

病情指数两次差分变换后的折线图（略），基本具有平稳走势的特征，其均值为 0.001 1，标准差为 0.010 1。对两次差分变换进行自相关函数（*ACF*）检验（图略），对于 $k>0$ 的各时滞，所有的“钉子”都在两条“约略估计”线以内，即所有时滞 k 的自相关系数 r_k 值都有 $|r_k|<\frac{2}{\sqrt{n}}$。对两次差分变换序列的各时滞的自相关系数 r_k 分别进行检验，$P>0.05$，即各时滞总体自相关系数 $\rho_k=0(k>0)$ 均成立。*Daniel* 检验法、修正的 *Box-Pierce Q* 检验法均证明两次差分序列具有平稳性。

2.2　模型的识别　对病情指数进行两次差分以后的序列进行模型识别。从样本自相关函数（*SACF*）、样本偏相关函数（*SPACF*）图（图略）可以看出 *AR*（*p*）是拖尾的，*MA*（*q*）是截尾的，初步判定模型为 *AR*（*p*）模型。

2.3　两次差分序列的模型估计和诊断过程　结合残差方差图定阶法、*F* 检验定阶法、最佳准则函数定阶法，对模型的阶数进行最后确定。

2.3.1　残差方差定阶法　假定模型是有限的自回归模型，如果选择的阶数 p 小于真正的阶数，则是一种不足拟合，因而剩余平方和 Q 必然偏大，残差方差 $\hat{\sigma}_\alpha^2$ 将比真正模型的残差方差 σ_α^2 大，这是因为我们把模型中本来应有的一些高阶项略去了，而这些项对于减少残差方差是有明显贡献的。另一方面，如果 p 已经达到真值，那么再进一步增加阶数，就是过度拟合，这并不会使 $\hat{\sigma}_\alpha^2$ 有显著减少，甚至还略有增加。这样用一系列阶数逐渐递增的模型进行拟合，每次都求出 $\hat{\sigma}_\alpha^2$，然后画出 p 和 $\hat{\sigma}_\alpha^2$ 的图形，即残差方差图。$\hat{\sigma}_\alpha^2$＝模型的剩余平方和/（实际观察值个数－模型的参数个数）。

用上述方法分别拟合 *AR*（1）至 *AR*（12）模型。首先通过参数估计值的检验，判断模型中是否保留平均项 *MU*，经多次试验，发现上述所有的 *AR* 模型，*MU* 项的 $|t|$ 值均很小，不能否定平均项为零的假设。计算各 *AR* 模型的 $\hat{\sigma}_\alpha^2$，然后划出残差方差图如图 1。

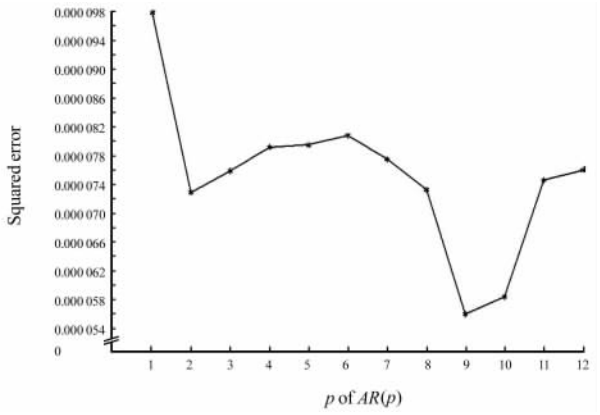

图 1　*AR*（*p*）残差方差图

Fig 1　Squared error of *AR*（*p*）

从图 1 可以看出自回归模型的残差方差有 3 个低点，阶数选择 9 或 10 时，残差方差分别为 0.000 056、0.000 058，要比阶数为 2 时的残差方差 $\hat{\sigma}_\alpha^2=0.000\ 073$ 小。因此初步确定，*AR*（*p*）模型的阶数可以在 2、9、10 这 3 个具有比较小的残差方法的阶数中选择。

2.3.2　*AR*（9）、*AR*（10）模型的参数估计　对初步确定的模型进行参数估计，所有的参数要经检验有统计学意义。SAS、SPSS 软件提供的是极大似然法及最小二乘法估计，并且提供参数的标准误，并对模型的系数是否为零进行 t 检验。公式为：t＝（估计的系数-假设检验估计的系数值）/（估计系数的标准误）。该统计量主要用来检验估计系数参数是否进入模型。

在 *AR*（*p*）模型中，当阶数为 9 时，模型参数的条件最小二乘估计中，*AR*1，2 至 *AR*1，7 的参数的 $|t|$ 值都过于小，P 值均大于 0.05，虽然对于小样本，该估计有一定的不可靠性。如果拟合只含有 *AR*1，1、*AR*1，8、*AR*1，9项的 *AR*（9）的模型，因为本研究采用3d采样1次，就意味着当前的病情指数

与前 24 d 至前 27 d 的病情指数有关,从霜霉病流行规律上看,不合适[2]。$AR(10)$同样存在这一现象。

2.3.3 AR(9)、AR(10)模型的白噪声检验 在从白噪声检验结果看(表略),在时滞为 6 时,独立性检验的 P 值小于 0.05,表明此时剩余项不独立,因此模型 $AR(9)$不合适。经试验,$AR(10)$也不成立。

2.3.4 $AR(2)$模型的参数估计、相关系数矩阵、白噪声检验和拟合优度检验 不保留平均项的$AR(2)$模型,参数的条件最小二乘估计如表 1,P 值均小于 0.05,参数可以进入模型。$|AR1,2| = -0.554\ 7 < 1$, $AR1,2+AR1,2<1$, $AR1,2-AR1,1<1$, 说明系数符合平稳性条件。

表 1 模型 $AR(2)$参数的条件最小二乘估计

Tab 1 Conditional least squares estimation for model $AR(2)$

Parameter	Estimate	Standard error	t	Approx Pr> |t|	Lag
AR1,1	−0.487 14	0.175 71	−2.77	0.010 4	1
AR1,2	−0.554 71	0.176 57	−3.14	0.004 3	2

各估计系数之间的相关系数不宜过大,否则意味着参数的质量不好。一般认为,若模型中任何两个被估计的参数的相关系数的绝对值大于或等于 0.9,则应怀疑模型结构是否稳定。从参数估计值的相关系数矩阵(表 2)可以看出,$AR1,1$ 和 $AR1,2$ 两个参数的相关系数为 0.296,远小于 0.9,不可能引起模型不稳定。

表 3 是白噪声检验的结果,各时滞点的剩余项独立性检验 P 值大于 0.05,表明剩余项独立,因此模型 $AR(2)$合适。

表 2 模型 $AR(2)$参数估计的相关系数矩阵

Tab 2 Correlation of parameter estimates for model $AR(2)$

Parameter	AR1,1	AR1,2
AR1,1	1.000	0.296
AR1,2	0.296	1.000

表 3 模型 $AR(2)$剩余项自相关性检验

Tab 3 Autocorrelation check of residuals for model $AR(2)$

To Lag	χ^2	DF	Pr> χ^2	Autocorrelation					
6	0.48	4	0.975 5	0.015	0.004	−0.066	0.035	0.038	0.080
12	7.63	10	0.664 9	−0.123	−0.351	−0.059	0.112	0.039	−0.113
18	13.15	16	0.661 6	0.181	−0.066	0.166	0.150	0.004	0.007
24	14.54	22	0.881 2	0.015	0.102	0.026	−0.002	−0.013	−0.017

最佳准则函数法是确定模型的优劣的常用方法。$AR(2)$模型的 AIC 值为−178.674 9,SBC 值为−176.083 2。

2.3.5 $AR(2)$模型的表达式 经过模型的识别、估计和诊断过程,得出了建立一个两次差分以后的 $AR(2)$模型是合适的,即 $ARIMA(2,2,0)$模型,其表达式为:$(1+0.481\ 7B+0.554\ 7B^2)(1-B)^2 y_t = a_t$。

2.4 预测过程 用 $ARIMA(2,2,0)$模型对黄瓜霜霉病序列进行预测,得残差方差 $\hat{\sigma}_a^2 = 0.000\ 073$, $RMSE = 0.008\ 537$, $SSE = 0.001\ 822$。$W = 0.951\ 1$, $P = 0.2283$,满足正态性。剩余项独立性检验表明残差独立。

从图 2 可看出,预测值和原始病情指数吻合得相当好,所有的实际值均在预测值 95%的可信区间内。残差图(略)显示,各时间点的残差均匀无趋势地散布在零轴两侧,显示模型建立效果良好。该模型是用前 29 个时间点的数据进行建模获得的,预留的第 30 个时间点的数据作为验证数据,其病情指数为 0.427 5,计算该点的预测值及可信区间、残差,对模型进行验证,模型的预测值为 0.420 5,标准误为 0.008 5,95%的可信区间为[0.403 7,0.437 2],与原始值吻合较好。

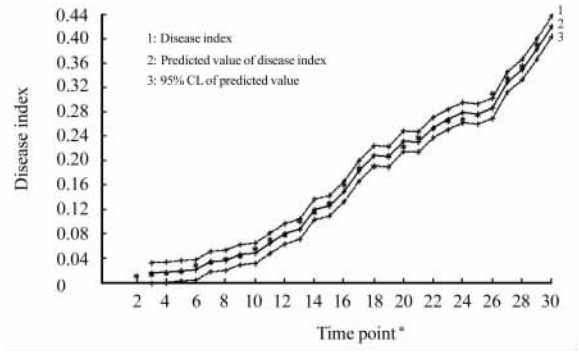

图 2　*AR*(2)预测效果图

Fig 2　Forecasting effect of *AR*(2)

*：Obseruation was carried out every 3 days

3　讨　论

时间序列分析模型参数估计的算法很多，相关矩估计是常用的方法（如 *Yule-Walker* 方程），但其精度有时偏低，有时针对具体的模型，求解相当复杂。为解决运算困难，对不同的模型可以采用不同的优化算法。为保证一定预测精度，本建模过程采用条件最小二乘估计。

在建模过程中，首先通过模型识别初步判定病情指数两次差分序列的模型为 $AR(p)$模型，然后用残差方差定阶法初步确定，$AR(p)$模型的阶数可以在 2、9、10 这 3 个具有比较小的残差方法的阶数中选择。再通过模型的参数估计、相关系数矩阵、白噪声检验和拟合优度检验确定最佳模型为 $AR(p)$模型，即 $ARIMA(2,2,0)$模型。上述通过残差方差定阶法缩小模型选择范围，再结合其他手段最后确定模型的思路，有利于快速准确找到合适的模型。

ARIMA 模型是经过序列平稳化、模型识别、参数估计和诊断、预测等严格的过程建立的，预测精度较高，但一般要求序列长度要达到 30，本建模过程使用 29 个时间点的数据，基本保证了模型的稳定性和预测的准确性。如果疾病数据序列长度足够，使用该种方法建模和预测是较好的选择。如果序列长度太短，就会造成模型不稳定，影响预测效果。如果按照 3 d 观察 1 次的采样频次，使用前两个月的数据建立 *ARIMA* 模型，对后面 1 个月的病情实施预测，在实践中是实用的。为了能够尽早建立预测效果的 *ARIMA* 模型，建议可以采样频次改为每天观察 1 次。

对本研究建立的 4 种一维时间序列模型：$ARIMA(2,2,0)$模型、单参数双重指数平滑模型、*Holt-Winters* 两参数双重指数平滑模型、基于传染病模型的自回归模型，从根均方误差、残差的正态性、自相关性、异方差、验证数据的预测误差以及残差图等方面进行了比较，发现 $ARIMA(2,2,0)$模型为一维时间序列分析方法中相对较理想的模型。该模型适合于霜霉病中期、后期的预测。在霜霉病发病早期，序列较短，此时使用 *ARIMA* 模型的条件还不具备，但预测工作要尽早开展，单参数双重指数平滑法、*Holt-Winters* 两参数双重指数平滑法在序列较短的情况下就可以建模，可以与本方法配合使用。

[参 考 文 献]

[1] 刘晓宏，金丕焕，陈启明. *ARIMA* 模型中时间序列平稳性的统计检验方法及应用[J]. 中国卫生统计，1998，15：14-16.

[2] 何自福，虞　皓，朱天圣，等. 广州地区黄瓜霜霉病流行速率的预测模型[J]. 植物保护，2001，27：10-12.

[收稿日期] 2006-06-30　　[修回日期] 2006-07-04

[本文编辑] 尹　荼