ARIMA 乘积季节模型在我国甲肝发病预测中的应用

王 超[1],丁 勇[2],陆 群[3],吴 静[2*]

([1]南京医科大学生物医学工程系,[2]数学与计算机教研室,[3]第一临床医学院,江苏 南京 210029)

[摘 要] **目的**:应用求和自回归移动平均模型(ARIMA)乘积季节模型对我国病毒性甲型肝炎进行预测分析,为甲型肝炎的防治提供决策依据。**方法**:对 1994~2012 年我国甲型肝炎月发病数的历史疫情数据建立 ARIMA 乘积季节模型,应用 Eviews 6.0 软件进行模型拟合,对 2013 年上半年甲型肝炎的月发病数进行预测,并用实际数据评估模型预测效果。**结果**:ARIMA$(1,1,0)(2,1,2)_{12}$模型较好地拟合了既往甲肝的实际发病序列,也获得了较好的预测效果。**结论**:ARIMA 模型能够较好地模拟我国甲型肝炎的发病趋势,预测效果良好,可为甲肝疫情的防控提供一定的科学数据。

[关键词] ARIMA 乘积季节模型;时间序列;甲肝;预测

[中图分类号] R512.6 [文献标志码] A [文章编号] 1007-4368(2014)01-075-05

doi:10.7655/NYDXBNS20140119

Application of multiple seasonal ARIMA model in forecasting the incidence of hepatitis A in China

Wang Chao[1], Ding Yong[2], Lu Qun[3], Wu Jing[2*]

(*[1]Department of Biomedical Engineering, [2]Department of Mathematics and Computer, [3]The First Clinical College, NJMU, Nanjing 210029, China*)

[Abstract] **Objective**: To forecast the incidence of hepatitis A in China by multiple seasonal autoregressive integrated moving average (ARIMA) model, and to provide decision basis for prevention and control of hepatitis A. **Methods**: ARIMA model was established according to the data of monthly reported hepatitis A cases in China from Jan.1994 to Dec.2012. Eviews 6.0 software was performed to construct the ARIMA model, and the constructed model was applied to predict monthly incidence in the first half of 2013. The model was evaluated by actual data. **Results**: ARIMA $(1,1,0)(2,1,2)_{12}$ exactly fitted the incidence of the previous time series, and got a good result on the predictive incidence in 2013. **Conclusion**: The multiple seasonal ARIMA model can be performed to make a short-term prediction and a dynamic analysis on the incidence of hepatitis A in China, and can provide a scientific basis for the prevention and control of hepatitis A.

[Key words] multiple seasonal ARIMA model; time series; hepatitis A; prediction

[Acta Univ Med Nanjing, 2014, 34(01): 075-079]

甲型病毒性肝炎,简称甲型肝炎、甲肝,是由甲型肝炎病毒(HAV)引起的、以肝脏炎症病变为主的传染病,主要通过粪-口途径传播。粪口传播的方式是多样的,一般情况下,日常生活接触传播是散发性发病的主要传播方式,因此在集体单位如托幼机构、学校和部队中甲型肝炎发病率较高。水和食物的传播,特别是水生贝类如毛蚶等是甲型肝炎暴发流行的主要传播方式。1988 年春季的上海甲肝大暴发,主要就是由市民食用受到 HAV 污染的毛蚶引起的,此次疫情波及面极广,短短 4 个月内罹患人数达到 31 万之多,造成了极大的社会恐慌。甲肝患病人群遍及各年龄段,主要为儿童和青少年。成人甲肝的临床症状一般较儿童重。冬春季常是甲肝发病的高峰期。自上海甲肝大暴发后,政府和卫生部门对甲肝的防控形势十分关注。20 世纪 90 年代初,国产甲肝减毒活疫苗问世,90 年代中期进口甲肝灭活疫苗进入中国市场,疫苗给甲肝的防控提供了有效的手段,同时伴随着社会经济的发展和生活习惯的改变等诸

[基金项目] 江苏省教育厅大学生实践创新训练计划项目(2012JSSPITP1033);江苏高校优势学科建设工程资助项目

*通信作者(Corresponding author),E-mail:wujing@njmu.edu.cn

多影响，甲肝的报告发病率呈现总体下降的趋势。但从目前官方公布的数据来看，相比其他的一些传染病，甲肝的发病人数相对较高，并且在一些地区，人群甲肝免疫力下降，极易导致甲肝的暴发和流行，因此甲肝仍是危害公众健康的重要传染病之一，其防控形势依然很严峻[1-2]。

求和自回归移动平均模型（autoregressive integrated moving average，ARIMA）是实现动态预测的模型工具，在多个学科和行业得到广泛应用[3]。该方法结合了自回归和移动平均方法的长处，具有不受数据类型束缚和适用性强的特点，且可以按年份、季度、月份进行预测。本文尝试用该模型对我国甲肝流行规律进行建模，并预测其未来的发病趋势，以期为早期发现甲肝的流行情况及制定相关的防治策略提供数据，从而为甲肝研究提供新的方法。目前尚未见利用该模型对我国甲肝发病趋势开展预测的文献报道。

1 资料和方法

1.1 资料

数据资料来源于我国卫生部网站（网址：http://www.moh.gov.cn）2004年1月~2013年6月的全国法定报告传染病疫情资料，其中2004年1月~2012年12月的数据用于建立模型，2013年1~6月数据用于验证模型的预测效果。

1.2 方法

1.2.1 ARIMA模型的基本原理

ARIMA模型是由美国统计学家Box和英国统计学家Jenkins于20世际70年代初提出的著名时间序列分析、预测和控制的方法，故又称Box-Jenkins建模方法。同时带有季节性与趋势性的ARIMA模型可以表现为相乘的形式，即ARIMA(p,d,q)(P,D,Q)s乘积季节模型。该模型有7个参数，其中p、q分别表示自相关函数（autocorrelations function，ACF）和偏自相关函数（partial autocorrelations function，PACF）的阶，d表示进行差分的次数；P、Q、D分别表示季节性自相关、偏自相关函数的阶和差分的次数，s表示季节性的周期。其一般形式为：

$$\Phi(L)U(L^S)\Delta^d\Delta^D{}_S Y_t=V(L^S)\Theta(L)\varepsilon_t \quad (1)$$

其中 $\Phi(L)=1-\emptyset_1L-\emptyset_2L^2-\cdots-\emptyset_pL^p$，$\Theta(L)=1-\theta_1L-\theta_2L^2-\cdots-\theta_qL^q$，$U(L^S)=1-u_1L^s-u_2L^{2S}-\cdots-U_PL^{PS}$，$V(L^S)=1-v_1L^S-v_2L^{2S}-\cdots-V_QL^{QS}$。$\varepsilon_t$ 代表独立扰动或随机误差。$\Phi(L)\Delta^dY_t$ 表示同一周期内不同周期点的相关关系，$U(L^S)\Delta^D{}_S$ 则描述了不同周期的同一周期点上的相关关系，二者结合便同时刻画了两个因素的作用[6-7]。当P=D=Q=0时，该模型便是一般的ARIMA模型。

1.2.2 建模过程

序列的平稳化：ARIMA建模的前提条件是时间序列的平稳性，对于非平稳序列，可通过数据变换和差分实现序列的平稳化。通常对存在异方差（方差不相同）的原序列进行自然对数变换，然后根据变换后序列的自相关和偏自相关图，确定非季节差分阶数d和季节差分阶数D，d和D宜取较低阶（一般不超过2）。s可以根据疾病的背景知识获得。

模型的识别：根据变换后平稳时间序列的自相关和偏自相关图，估计模型p、q、P、Q的值。

参数估计及检验：一般运用最大似然法（ML）或无约束最小二乘法（ULS），估计模型的系数，并对其显著性进行检验。

模型的诊断检验：一个合适的模型的残差序列应是白噪声序列。得到白噪声序列，就说明时间序列中有用的信息已经被提取完毕了，剩下的全是随机扰动，如果残差不是白噪声，就说明残差中还有有用的信息，需要修改模型或者进一步提取该信息。可应用Box-Ljung Q统计量进行检验，其ACF和PACF不应与0有显著性差异。

模型的预测：利用所建模型进行预测，从而评价模型的优劣，也是其实际价值的体现。

1.2.3 实现软件

采用Eviews6.0软件进行数据的处理和分析[8-9]。

2 结 果

2.1 序列的平稳化

图1是2004年1月~2012年12月我国甲型肝炎月发病数（Yt）时间序列图，从图1可以看出，该序列呈现出明显的非平稳性和季节性（s = 12），并随着时间呈现递减型异方差。甲肝发病数显现总体下降趋势，有明显的季节效应，一般在每年的3月份前后常会出现病例数的高峰。为消除异方差，本研究首先对原始数据进行自然对数转换，以平稳序列的方差。从对数变换后的甲肝月发病数据（LnYt）的自相关图和偏自相关图（图2）可以看出，自相关图衰减很慢，LnYt是非平稳的，且相关图存在周期为12个月的季节波动。因此对LnYt进行一阶非季节差分和一阶季节差分，得到新的序列 $\triangle\triangle_{12}$LnYt。从序列 $\triangle\triangle_{12}$LnYt 的相关图和偏相关图（图3）可以看出，其自相关函数值快速衰减，近似为一个平稳过程。选择不含常数项和趋势项的ADF检验，检验结果

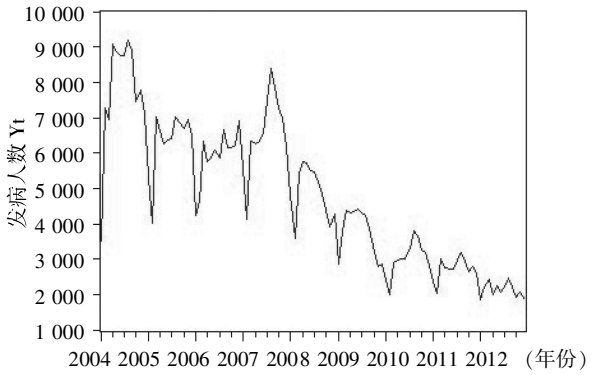

图 1　我国甲型肝炎月发病数(Yt)时间序列图

Figure 1　Time series of monthly incidence of hepatitis A (Yt) in China

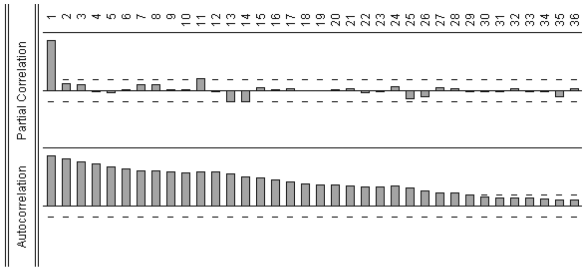

图 2　序列 LnYt 的相关图(下)和偏相关图(上)

Figure 2　Autocorrelation (below) and partial correlation (above) of series LnYt

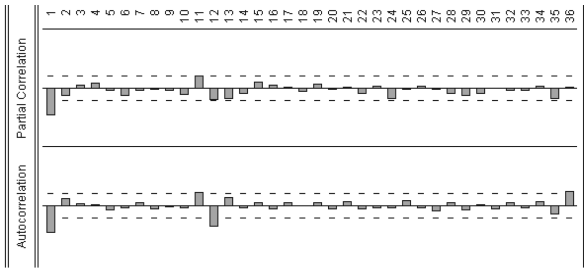

图 3　序列 $\triangle\triangle_{12}$LnYt 的相关图(下)和偏相关图(上)

Figure 3　Autocorrelation (below) and partial correlation (above) of series $\triangle\triangle_{12}$LnYt

也显示数据已呈现平稳性。此时数据的曲线见图 4。

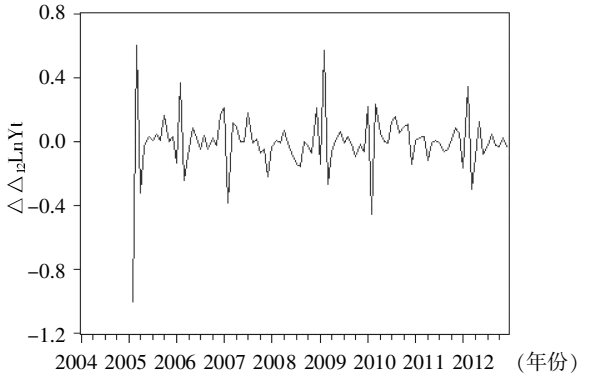

图 4　对数变换一阶差分、一阶季节差分后序列图

Figure 4　Series after natural log transformed and one time seasonal differenced

2.2　模型的识别与定阶

由于原始序列 Yt 对数变换后，经过一阶非季节差分和一阶季节差分达到平稳，因此 s = 12,d = 1,D = 1,模型可初步确定为 ARIMA(p,1,q)(P,1,Q)$_{12}$,其中 p、q、P 和 Q 是待定的参数。观察序列 $\triangle\triangle_{12}$LnYt 的偏相关图，序列$\triangle\triangle_{12}$LnYt 的偏相关函数在滞后 1 阶以后降为零,因此 p = 1;自相关函数在滞后 1 阶以后降为零,因此 q = 1。参数 P、Q 的判断较为困难,但根据文献,参数超过 2 阶的情况很少见,可以分别取 2,1,0 由高阶到低阶逐个试验,根据模型的拟合优度、残差情况以及系数间的相关性进行综合判断,以确定这两个参数[6-7]。

2.3　参数估计及检验

本研究对 24 个模型用 Eviews6.0 软件进行了计算,其中 2 个备选模型的参数具有统计学意义,结果见表 1。从该表可看出,ARIMA(1,1,0)(2,1,2)$_{12}$ 和 ARIMA (1,1,1)(2,1,1)$_{12}$ 模型的所有参数都通过了 t 检验。同时模型所有根(包括实根和复根)的

表 1　备选模型的参数估计结果

Table 1　Parameter estimates of alternative ARIMA models

参数	ARIMA(1,1,0)(2,1,2)$_{12}$			ARIMA(1,1,1)(2,1,1)$_{12}$		
	系数	t 值	P 值	系数	t 值	P 值
AR(1)	−0.250	−2.672	0.010	−0.562	−3.747	<0.001
MA(1)				0.467	2.469	0.016
SAR(12)	−0.628	−6.644	<0.001	−0.991	−14.435	<0.001
SAR(24)	−0.552	−6.394	<0.001	−0.800	−11.665	<0.001
MA(12)	0.053	1.889	0.063	0.888	−35.249	<0.001
MA(24)	−0.881	−32.800	<0.001			
R^2 值		0.830			0.751	
调整 R^2 值		0.819			0.736	
AIC 统计量		−2.728			−2.348	

倒数均小于1，均符合建模要求。但从拟合优度来看，ARIMA$(1,1,0)(2,1,2)_{12}$模型的R^2值和调整R^2值均比ARIMA$(1,1,1)(2,1,1)_{12}$的大，而对于AIC统计量，ARIMA $(1,1,0)(2,1,2)_{12}$模型比ARIMA $(1,1,1)(2,1,1)_{12}$的小，因此，ARIMA $(1,1,0)(2,1,2)_{12}$模型的拟合效果较好，根据表1估计的参数值，得到公式（1）的表达式为：$(1+0.250L)(1+0.628L^{12}+0.552L^{24})(1-L)(1-L^{12})\ln(Yt)=(1+0.053L^{12}-0.881L^{24})\varepsilon_t$。

2.4　模型的诊断

在Eviews6.0中对ARIMA$(1,1,0)(2,1,2)_{12}$模型选择残差检验的Q统计量检验，结果见表2。根据各滞后期Q统计量的P值，检验结果不能拒绝残差不相关的零假设，即ARIMA$(1,1,0)(2,1,2)_{12}$模型的残差序列是白噪声序列，从而提示所选模型恰当，用于预测是合适的。

表2　残差序列的自相关检验

Table 2　Autocorrelation check of residuals

滞后阶数	Q统计量	P值	自相关函数值
6	5.316	0.021	0.062,0.033,0.149,-0.126,0.086,0.138
12	8.744	0.272	-0.139,-0.041,-0.009,-0.078,-0.050,0.109
18	12.152	0.515	0.017,0.050, -0.014,0.008,-0.156,-0.092
24	19.662	0.415	-0.112,-0.205,0.020,0.020,-0.132,-0.030

2.5　模型的拟合与预测

本研究用ARIMA $(1,1,0)(2,1,2)_{12}$模型对原序列进行了拟合（图5），拟合值的动态趋势表现出与实际值极为相似的升降规律，较好地模拟出时间序列的波动趋势和季节要素。

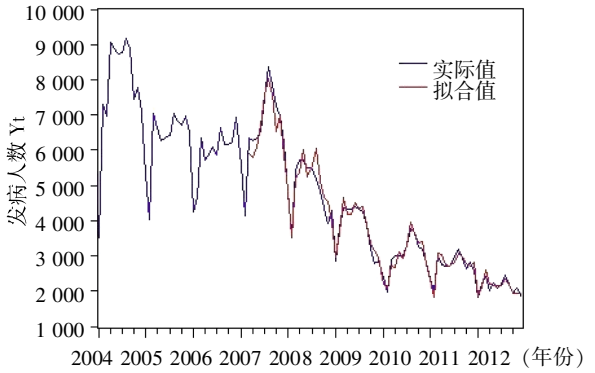

图5　实际值与模型拟合值序列比较图

Figure 5　Actual value and fitted value series

用ARIMA $(1,1,0)(2,1,2)_{12}$模型预测我国2013年1~6月甲肝逐月发病人数，并结合卫生部网站公布的实际数据进行预测精度的验证，结果见表3，该模型的预测结果相对误差率最大为9.5%，而经计算，另一个模型ARIMA$(1,1,1)(2,1,1)_{12}$的预测结果相对误差率最大为12.4%，显然ARIMA$(1,1,0)(2,1,2)_{12}$模型的预测效果较好，这与前面拟合优度检验的结果是一致的。以上结果说明模型的选择是正确的，可将其运用于甲肝发病情况的分析与预测。

表3　ARIMA$(1,1,0)(2,1,2)_{12}$模型的预测结果

Table 3　Results of monthly incidence of hepatitis A (Yt) by ARIMA$(1,1,0)(2,1,2)_{12}$

月份	实际值	预测值	绝对误差	相对误差
1	1 610	1 456.797	153.203	0.095
2	1 403	1 281.448	121.552	0.087
3	1 951	1 775.659	175.341	0.090
4	1 830	1 757.950	72.050	0.039
5	1 908	1 850.698	57.302	0.030
6	1 666	1 780.993	114.993	0.069

3　讨　论

目前国内对甲肝的预测研究报告不多，且基本都是对甲肝的地区年发病情况进行分析预测，未见对全国的发病情况进行逐月预测研究。对于其他传染病预测研究方面，大致有包括回归模型、微分方程模模型、灰色预测模型在内的非周期模型和包括余弦模型、Markov模型在内的周期模型[10]。回归模型对变量的分布有特殊要求，且要求变量独立，无法考虑到预测变量的自相关性；微分方程模型描述的是传染病的自然发展过程，没有考虑人会采取有效的措施进行控制、隔离或者治疗，得到结果也不理想；灰色预测模型进行预测，只需很少数据量即可预测，但当数据波动比较大时，预测精确度大大下降；余弦模型是研究周期现象的简单模型，不能对复杂序列进行准确分析；Markov模型对选定区间进行预测，预测准确度依赖人为设定的区间，精确度也不高。而ARIMA模型比较灵活，既适用于非周期性序列，也适用于周期性序列，且周期可以为年份、季度、月份，故适用范围更广泛。近年来，该模型已广泛应用于社会、经济、医疗卫生等多个领域[11-12]，特别在传染病的预测方面，是一种有效而实用的方法[13-14]。

由历史数据看出,甲型肝炎的月发病数序列不仅含有时间趋势性成分,还混有季节性成分。如果只是考虑单一的因素建立模型就不能做到较准确地对其进行预测。本文运用的 ARIMA 乘积季节模型整合了趋势因素、周期因素和随机误差等因素的原始时间序列变量,通过差分数据转换等方法将非平稳序列转变为零均值的平稳随机序列,通过反复识别和模型诊断比较并选择理想的模型进行数据拟合和预测,是一种短期预测效果较好的模型。本文首次用该模型研究我国甲肝的月发病情况,并取得较好的效果,同其他甲肝发病预测的文章相比,相对误差明显减小[15]。

但是该模型也有不足之处:①建立此模型需要一定数量的历史数据;②所建立的模型只能用于短期预测,当获得新数据时,应不断加入新的实际值,以修正或重新拟合更优的模型。当实际问题比较复杂时,模型的建立比较困难,应用者需要有清晰的思路、对实际资料有深刻的理解,并不断积累经验,寻找对序列进行平稳性处理的方法以提高预测模型的精度,这样才能获得更合适的模型,得到更好的效果。

[参考文献]

[1] Donnan EJ,Fielding JE,Gregory JE,et al. A multistate outbreak of hepatitis a associated with semidried tomatoes in Australia,2009[J]. Clin Infect Dis,2012,54(6):775-781

[2] Gallot C,Grout L,Roque-Afonso AM,et al. Hepatitis A associated with semidried tomatoes,France,2010 [J]. Emerg Infect Dis,2011,17(3):566-567

[3] Bowerman BL,TO'Connell R. Forecasting and time series:an applied approach [M]. 3rd ed. China Machine Press,2003:437-595

[4] 胡建利,祖荣强,彭志行,等. 江苏省戊型肝炎发病趋势的时间序列模型应用[J]. 南京医科大学学报:自然科学版,2011,31(12):1874-1878

[5] 丁晓艳,彭志行,陶　红,等. 运用时间序列模型对麻疹流行趋势的预测与分析[J]. 南京医科大学学报:自然科学版,2011,31(8):1200-1202

[6] 方积乾,陆　盈. 现代医学统计学[M]. 北京:人民卫生出版社,2002:219-269

[7] Box GEP,Jenkins GM. Time series analysis forecasting and control[M]. Holden-Day,1976

[8] 攸　频,张晓峒. Eviews 6 实用教程[M]. 北京:中国财政经济出版社,2008:117-150

[9] 易丹辉. 数据分析与 EVIEWS 应用[M]. 北京:中国人民大学出版社,2008:137-140

[10] 王丙刚,曲　波,郭海强,等. 传染病预测的数学模型研究[J]. 中国卫生统计,2007,24(5):536-540

[11] 吴　静,丁　勇,刘成友. 甲型 H1N1 流感病毒 HA 基因的简单重复序列预测[J].南京医科大学学报:自然科学版,2013,33(1):42-47

[12] 赵　凌,张　健,陈　涛. 基于 ARIMA 的乘积季节模型在城市供水量预测中的应用 [J]. 水资源与水工程学报,2011,22(1):58-62

[13] 吴孟泉,赵　凯. 基于 ARIMA 模型的 2009 年山东省手足口病疫情分析及预测 [J]. 鲁东大学学报: 自然科学版,11,27(1):71-75

[14] 彭志行,鲍昌俊,赵　杨,等. ARIMA 乘积季节模型及其在传染病发病预测中的应用 [J]. 数理统计与管理,2008,27(2):362-368

[15] 朱奕奕,赵　琦,冯　玮,等. 应用指数平滑法预测上海市甲型病毒性肝炎发病趋势 [J]. 中国卫生统计,2013,30(1):31-36

[收稿日期] 2013-07-06