·公共卫生与预防医学研究·

ARIMA乘积季节模型预测我国戊肝的发病趋势

丁 勇[1],吴 静[2],武 丹[1],李 婉[1],张蓓蓓[1*]

[1]南京医科大学康达学院,江苏 连云港 222000;[2]南京医科大学生物医学工程与信息学院,江苏 南京 210029

[摘 要] **目的**:根据戊型肝炎(戊肝)季节性、趋势性流行现象,建立求和自回归移动平均(autoregressive integrated moving average,ARIMA)乘积季节模型对我国戊肝发病进行预测。**方法**:应用SPSS23.0软件对2004年1月—2018年6月我国戊肝传染病疫情月度数据建模,对2018年下半年戊肝发病数进行预测,以该时段疫情数据评估模型的预测效果。**结果**:将ARIMA(2,1,0)$(0,1,1)_{12}$和ARIMA(0,1,2)$(0,1,1)_{12}$两个模型预测的平均值作为预测值,预测结果的平均相对误差为4.69%,标准差为3.27%。**结论**:ARIMA乘积季节模型拟合及预测效果良好,能够较好地描述该时段我国戊肝的发病趋势,为戊肝预防控制措施的制定以及卫生资源的合理配置提供一定的科学依据。

[关键字] 戊型肝炎;ARIMA乘积季节模型;时间序列;预测

[中图分类号] R183.1 [文献标志码] A [文章编号] 1007-4368(2020)11-1725-05

doi:10.7655/NYDXBNS20201128

The incidence of hepatitis E in China predicted by multiple seasonal ARIMA model

DING Yong[1], WU Jing[2], WU Dan[1], LI Wan[1], ZHANG Beibei[1*]

[1]Kangda College, Nanjing Medical University, Lianyungang 222000; [2]School of Biomedical Engineering and Information, Nanjing Medical University, Nanjing 210029, China

[**Abstract**] **Objective**: According to the seasonal and trend epidemic phenomenon of hepatitis E, the multiple seasonal ARIMA model was established to predict the infectious incidence of hepatitis E in China. **Methods**: SPSS23.0 software was used to model the monthly data of the epidemic of hepatitis E infectious diseases in China from January 2004 to June 2018, so as to predict the incidence of hepatitis E in the second half of 2018 and to evaluate the prediction effect of the model through the epidemic data during the period. **Results**: The average values of the prediction of the two models, ARIMA(2,1,0)$(0,1,1)_{12}$ and ARIMA(0,1,2)$(0,1,1)_{12}$, were used as the prediction values, the average relative error of the prediction was 4.69% and the standard deviation was 3.27%. **Conclusion**: Results of fitting and prediction of the multiple seasonal ARIMA model are good. The model can better describe the incidence trend of hepatitis E in China during the period, and provide certain scientific basis for the formulation of preventive control measures against hepatitis E and reasonable allocation of health resources.

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

[J Nanjing Med Univ, 2020, 40(11): 1725-1729]

戊型病毒性肝炎(简称戊肝)由戊肝病毒(hepatitis E virus,HEV)引起,以肝脏炎症和坏死病变为主的急性病毒性肝炎,其临床症状和流行病学特征与甲型肝炎相似,在急性病毒性肝炎的死亡率中占首位[1]。普通人群感染后的病死率为0.5%~3.0%,妊娠女性HEV感染的年发生率为4.6%~5.6%,病死率高达10%~25%,尤其在妊娠后3个月的女性患者中病死率可达10%~39%[2]。HEV主要经粪-口途径传播,具有明显季节性特征,多见于雨季或洪水之后[3]。

戊肝的发病呈现世界性分布,主要见于亚洲和非洲的一些发展中国家,一般在发达国家以散发病例为主,发展中国家以流行为主[4-5]。我国是戊肝的

[基金项目] 南京医科大学科技发展基金(2017NJMU229);江苏省高校自然科学研究(19KJD330001)

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

主要流行地区之一，其防控工作一直是我国一项重要的公共卫生问题，近年戊肝发病率增高使得这一问题更为凸显。而现有的“疾病监测信息管理系统”缺少有效的预测预警机制，限制了戊肝预防控制工作的开展。基于统计分析和数学模型等方法对戊肝疫情发展规律进行预测，是戊肝疫情的控制、预防以及卫生决策过程中不可或缺的科学依据。求和自回归移动平均模型（autoregressive integrated moving average，ARIMA）是Box和Jenkins于70年代初提出的时间序列预测方法，故又称Box-Jenkins模型，是一种适用于平稳性时间序列的预测模型[6-7]。本文用ARIMA模型对我国戊肝疫情的时间序列进行分析和预测。

1　资料和方法

1.1　资料

数据资料来源于中国疾病预防控制局发布的2004年1月—2018年12月的全国法定传染病疫情报告[8]，其中2004年1月—2018年6月的戊肝疫情数据用于建立时间序列模型，2018年7月—2018年12月的数据用于检验模型预测的效果。

1.2　方法

在时间序列分解的基础上，对于带有季节周期性的时间序列，要采用考虑季节性的乘积模型ARIMA(p,d,q)(P,D,Q)s，其中参数p、q和d表示自相关函数（autocorrelations function，ACF）、偏自相关函数（partial autocorrelations function，PACF）的阶和差分的次数；参数P、Q和D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[9]。ARIMA模型要求时间序列为同方差的平稳序列。

本文采用EXCEL 2010软件对戊肝疫情数据进行汇总整理，通过统计分析软件SPSS 23.0对数据进行分析和处理，并按如下4个阶段建立ARIMA模型。①序列平稳化：对异方差的非平稳序列，通过对数（ln）转化与适当差分（▽），转化为平稳时间序列；②模型识别与定阶：采用Box-Jenkins方法，参考预处理后的平稳时间序列的ACF和PACF图，估计模型的p、q、P、Q的值；③参数估计及诊断：运用最大似然法估计ARIMA模型的系数，并对其显著性进行检验。采用白噪声残差检验对模型进行诊断[10-11]，再结合拟合系数R^2、贝叶斯信息量（Bayesian information criterion，BIC）最小信息准侧，建立适合的模型；④模型的预测：利用建立的模型进行预测，并结合实际数据对预测效果进行评估。

2　结　果

2.1　流行特征分析及序列平稳化

图1是我国2004年1月—2018年6月的戊肝发病人数时间序列图，显示我国戊肝发病人数呈现明显的上升趋势和季节性效应（按月划分，s=12）。为此，对时间序列进行季节性分解，时间序列的分解旨在将时间序列中的趋势、季节和不规则成分（随机误差）分离出来，分别进行统计分析[12]。时间序列分解得到的发病数季节因子如图2所示，戊肝发病的流行特性具有周期性的统计规律：从每年的10月份发病人数开始逐渐增加，到次年的3月份达到峰值，以后逐渐下降，2—4月为发病高峰期。

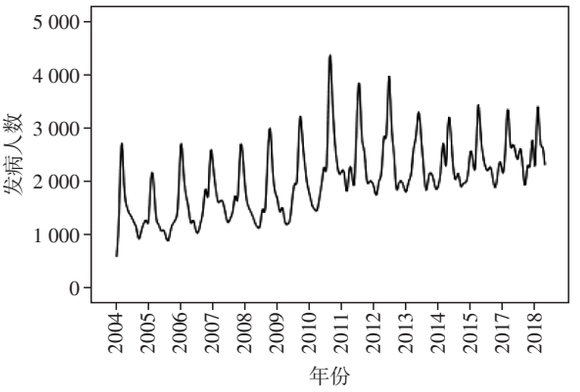

图1　我国戊肝发病人数时间序列图

Figure 1　Time series of hepatitis E incidence in China

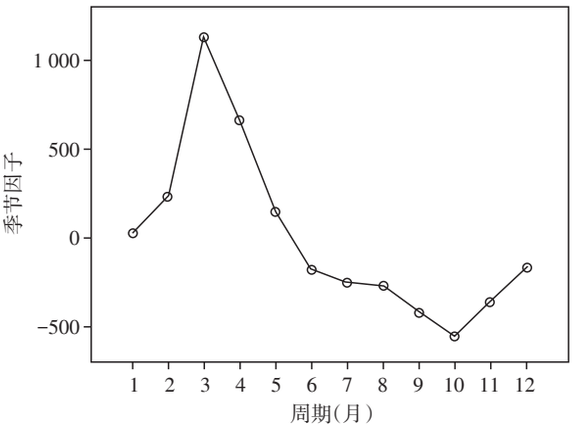

图2　戊肝发病数的季节因子

Figure 2　Seasonal factors of hepatitis E incidence

图1数据序列具有明显的异方差和非平稳特征，需要进行平稳性预处理：对序列进行自然对数转化以减小异方差，同时通过一阶差分（d=1）和一阶季节差分（D=1），消除序列的趋势性和季节性影

响,得到时间序列图(图3),序列在0附近呈现平稳的小幅上下波动,序列基本平稳。

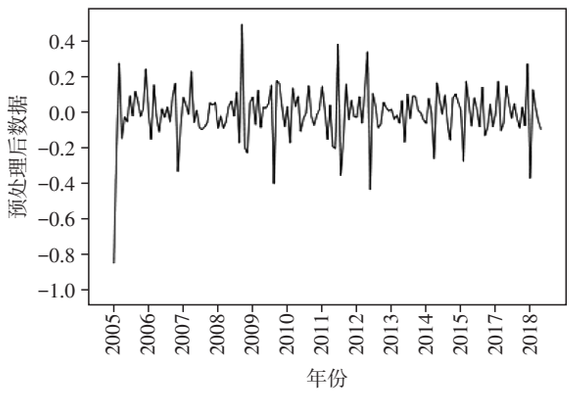

图3 预处理后数据时间序列图

Figure 3 Time series after preprocessing

2.2 模型的识别和定阶

由数据序列分解和平稳化过程可知,原始数据是以s=12个月为周期的季节性时间序列。经过自然对数转化并一阶差分后,原始序列的异方差和趋势性基本消失,可确定模型的参数d=1;经过一阶季节差分后,数据的季节性基本消失,可确定模型的参数D=1,故初步确定模型的基本形式为ARIMA$(p,1,q)(P,1,Q)_{12}$。为在较大范围内选择最佳模型,结合预处理后序列的ACF(图4)和PACF分析(图5),初步观察分析后取p=0、1、2和q=0、1、2进行筛选。参数P、Q的取值判定较为困难,根据已有相关研究成果,P、Q取值超过2阶的情况比较少见,故都取为0、1、2进行筛选[13]。这4个参数的不同选择共有3^4=81种备选模型,为确定这4个参数的最优取值,考虑模型的拟合效果、最小信息准则和残差序列等有关指标综合进行评价。

2.3 参数估计及诊断

采用SPSS 23.0中时间序列预测模块,分别对81个备选模型进行计算。通过杨-博克斯统计量的显著性要求和模型系数的T检验要求,剔除不满足要求的模型,得到如表1的两个备选模型。两个模型的系数均通过t检验($P<0.05$);从拟合优度方面来看,两个模型在拟合系数R^2和BIC值非常接近。表2为两个模型的残差是否为白噪声的Q统计量检验,各滞后阶数下Q统计量的P值均大于0.05,故两个备选模型的残差序列均为白噪声序列[14]。由上可知,两个备选模型均具有统计学意义,满足ARIMA建模的要求。

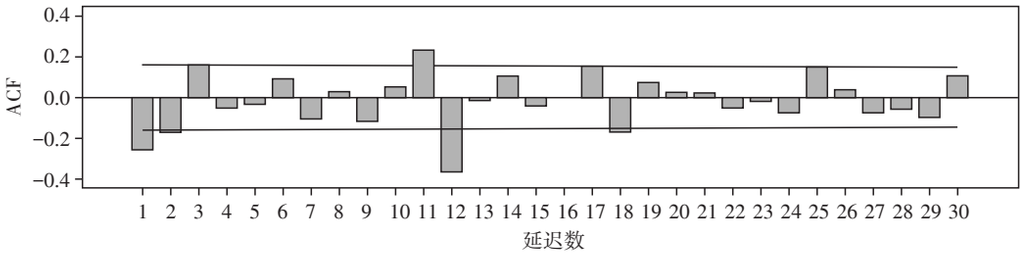

图4 预处理后序列自相关图

Figure 4 Autocorrelation of series after preprocessing

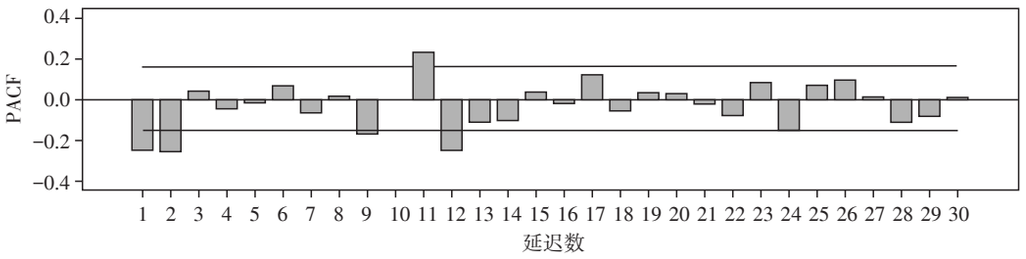

图5 预处理后序列偏自相关图

Figure 5 Partial autocorrelation of series after preprocessing

图6为两个备选模型的拟合曲线图,从直观效果来看,二者与实测值的变化规律及数值拟合上均有良好效果,能够较好地模拟出原始时间序列的波动规律和季节特性。

预测往往有误差,一般认为误差服从正态分布$N(\mu,\sigma^2)$。由统计知识可知,当样本数据服从正态分布时,样本均数服从正态分布$N(\mu,\sigma^2/n)$,均数减少了数据的波动性(方差)。所以保留这两个模型,并将它们预测的平均值作为实际的预测值。

表1　ARIMA模型参数估计检验及拟合结果统计表

Table 1　Statistics of parameter estimation test and fitting results of ARIMA models

变量	ARIMA $(2,1,0)(0,1,1)_{12}$			ARIMA $(0,1,2)(0,1,1)_{12}$		
	系数	t值	P值	系数	t值	P值
AR(1)	−0.343	−4.415	<0.001	—	—	—
AR(2)	−0.312	−4.392	<0.001	—	—	—
MA(1)	—	—	—	0.391	5.065	<0.001
MA(2)	—	—	—	0.358	4.652	<0.001
SMA(12)	0.694	8.567	<0.001	0.739	8.994	<0.001
常数	−0.012	−2.181	0.031	−0.011	−2.291	0.023
P值		0.362			0.232	
R^2		0.795			0.794	
BIC		11.522			11.525	

表2　ARIMA模型残差白噪声检验汇总表

Table 2　Residual white noise test of ARIMA models

滞后阶数	ARIMA $(2,1,0)(0,1,1)_{12}$		ARIMA $(0,1,2)(0,1,1)_{12}$	
	Q统计值	P值	Q统计值	P值
6	3.653	0.724	8.958	0.176
12	13.461	0.336	15.681	0.206
18	16.308	0.571	18.606	0.416
24	19.077	0.748	21.013	0.638

2.4　模型的预测和比较

用ARIMA$(2,1,0)(0,1,1)_{12}$模型进行预测称为方法1，用ARIMA$(0,1,2)(0,1,1)_{12}$模型进行预测称为方法2，将这两个模型预测值的平均值作为预测值称为方法3。

从表3可知，将预测的相对误差的平均值和标准差作为预测效果精度和稳定性的指标，从相对误差的平均值和标准差来看，ARIMA$(2,1,0)(0,1,1)_{12}$模型预测精度不如ARIMA$(0,1,2)(0,1,1)_{12}$模型，但稳定性较好。综合来看，将两个模型预测的平均值作为预测结果的方法3，具有保持预测精度且预测稳定的优点，提高了预测的可信度。

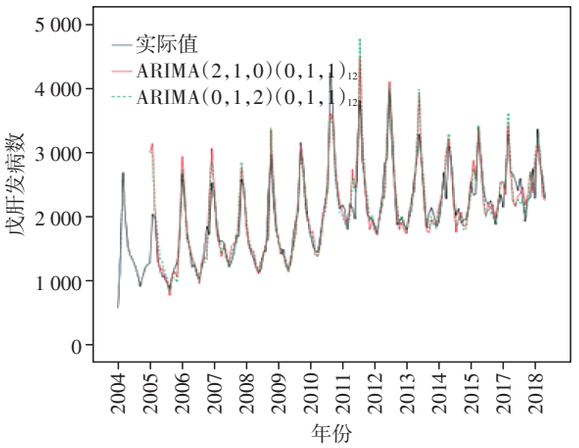

图6　模型拟合结果

Figure 6　Model fitting results

3　讨　论

传染病不仅使患者健康受到损害，危及生命，也造成家庭医疗费用负担。传染病暴发不仅削弱社会生产力，降低人均寿命，也带来医疗、卫生资源的极大损耗。健康中国是国家的战略目标，防控是最经济最有效的健康策略。传染病防控是公共卫生事业中的重要工作之一，关系到一个国家、地区人民的公共健康。传染病的分析和预测是对疾病未来发生、发展及流行趋势认知的重要手段，是制定传染病防控策略的重要前提[15]。要贯彻预防为主的卫生与健康工作方针，以防控“费”，将预防关口前移，避免小病酿成大疫。防控的关键措施之一是对传染病的准确预测。

建立合理的统计模型是进行预测研究的基础，目前使用较多的模型有微分方程模型、灰色预测模

表3　ARIMA模型预测结果比较

Table 3　Comparison of prediction results of ARIMA models

月份	实际戊肝发病数[a]	方法1		方法2		方法3	
		预测值(例)	相对误差(%)	预测值(例)	相对误差(%)	预测值(例)	相对误差(%)
7	2 386	2 208.0	7.46	2 181.0	8.59	2 194.5	8.03
8	2 368	2 288.0	3.38	2 220.0	6.25	2 254.0	4.81
9	2 023	2 131.0	5.34	2 070.0	2.32	2 100.5	3.83
10	1 896	1 965.0	3.64	1 930.0	1.79	1 947.5	2.72
11	2 264	2 285.0	0.93	2 245.0	0.84	2 265.0	0.04
12	2 335	2 549.0	9.17	2 527.0	8.22	2 538.0	8.69

方法1相对误差为(4.98 ± 2.99)%，方法2相对误差为(4.67 ± 3.43)%，方法3相对误差为(4.69 ± 3.27)%。a：我国2018年7月—2018年12月实际的戊肝发病数。

型、Markov模型、通径分析模型、ARIMA模型等,模型选取主要依赖于数据的自身特性[10]。与其他模型相比,ARIMA模型法结合了自回归和移动平均方法的优点,对各类型时间序列数据具有较强的适用性和灵活性,近年来在经济、社会、医学、信息学等领域中得到了广泛应用[16]。戊肝的流行是多种复杂因素综合影响的结果,其时间序列呈现趋势性、季节性和非线性的特征。由于不同时间的发病数不同,数据呈波动状态,形成异方差;每年的发病数有一定的自身规律,形成季节性的周期;传染病蔓延得到有效控制时,发病数呈下降的长期趋势。这些特点非常适合用ARIMA模型来描述其规律性,并进行预测。本文结合时间序列分解对样本数据进行了拟合和预测研究,结果表明ARIMA模型的拟合效果较好,预测精确较高,说明时间序列数据建立的ARIMA模型对传染病预测有推广价值。

目前国内对全国范围戊肝疫情分析预测报道较少,多为区域性传染病预测探索[5,11]。本文对我国戊肝月度发病数据,综合考虑季节性、趋势性等因素,建立ARIMA乘积季节预测模型,对全国范围内戊肝未来的流行趋势进行预测。用筛选的两个模型预测值的平均值作为预测结果,提高了预测的精度和可信度。时间序列分解显示,我国戊肝疫情爆发高点在2—4月,建议防控措施在2月前启动,做好防控宣传教育工作,建立针对孕妇、慢性肝炎患者、中老年等敏感群体的有效监控机制和应急预案,防止疫情扩散。从图1可知,最近几年,我国戊肝的防控取得了成效,发病数得到了有效控制,呈现平稳趋势。

本研究基于数据序列分析,需要历史数据的累积和新数据的及时补充,有一定的局限性,这对疾病监测信息系统的管理和完善也提出了要求[17]。另外在戊肝疫情防治中,精度好、可信度高的短期预测和趋势性的长期预测同样重要,因此探究融合多种方法的戊肝类传染病预测研究是进一步制定防控措施和合理配置卫生资源的新方向。

[参考文献]

[1] BALAYAN M S, ANDJAPARIDZE A G, SAVINSKAYA S S, et al. Evidence for a virus in non-A, non-B hepatitis transmitted via the fecaloral route[J]. Intervirology, 1983, 20(1):23-31

[2] 王晶晶,田德英. 戊型肝炎的研究现状和展望[J]. 临床肝胆病杂志,2013,29(2):84-87

[3] 马 涛,谭照营,祖荣强,等. 2005—2014年江苏省戊型病毒性肝炎流行病学特征分析[J]. 现代预防医学,2016,43(13):2310-2314

[4] DENNER J, PISCHKE S, STEINMANN E, et al. Why all blood donations should be tested for hepatitis E virus (HEV)[J]. BMC infectious diseases, 2019(1):541-543

[5] 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

[6] SHADAB A, SAID S S. Box-Jenkins multiplicative ARIMA modeling for prediction of solar radiation: a case study [J]. Inter J Ene Water Res, 2019, 3(4):305-318

[7] 张文娟,刘文东,胡建利,等. 基于ARIMA模型的江苏省梅毒疫情预测[J]. 南京医科大学学报(自然科学版),2017,37(5):649-652

[8] 中国疾病预防控制局,全国法定传染病疫情报告[EB/OL]. [2019-08-09]. http://www.nhc.gov.cn/jkj/new_index.shtml,2004.01-2018.12

[9] 罗兴甸,戴家佳,罗登菊. ARIMA乘积季节模型在我国麻疹发病预测中的应用[J]. 贵州大学学报(自然科学版),2019,36(3):9-14

[10] 王 超,丁 勇,陆 群,等. ARIMA乘积季节模型在我国甲肝发病预测中的应用[J]. 南京医科大学学报(自然科学版),2014,31(1):75-79

[11] 夏建华,张爱红,张红星,等. ARIMA乘积季节模型在如东县戊型肝炎发病预测中的应用[J]. 中国预防医学杂志,2016,17(2):120-123

[12] 邓维斌,周玉敏,刘 进,等. SPSS23统计分析实用教程[M]. 北京:电子工业出版社,2017:269-271

[13] 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

[14] 张生奎,王镇德,杨 荔,等. 基于SARIMA-ERNN组合模型预测我国细菌性痢疾发病率[J]. 南京医科大学学报(自然科学版),2019,39(6):925-931

[15] NAKATANI H. Global strategies for the prevention and control of infectious diseases and non-communicable diseases[J]. J Epidemiol, 2016(4):171-178

[16] ÜMIT Ç, BÜYÜK S, ŞEYDA E. Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition [J]. Neurocomputing, 2019, 5(99):151-163

[17] 祝丙华,王立贵,孙岩松,等. 基于大数据传染病监测预警研究进展[J]. 中国公共卫生,2016,32(9):1276-1279

[收稿日期] 2019-08-09