·论　著·

ARIMA模型在深圳市病毒性肝炎发病趋势预测的应用

郑慧敏*,薛允莲,黄燕飞,戴传文,姜世强

深圳市南山区疾病预防控制中心,广东 深圳 518054

摘要:目的　通过探讨单纯求和自回归滑动平均模型(ARIMA模型)应用于病毒性肝炎发病率预测的可行性,为当前的防控工作提供科学依据。**方法**　采用SAS9.2软件对深圳市2004~2013年的病毒性肝炎的月发病率进行ARIMA模型的建模拟合,预测2014年病毒性肝炎的月发病率,利用预测值和实际值的均方误差、平均绝对误差、平均绝对百分误差评价拟合效果,选择合适的模型预测深圳市2015年病毒性肝炎的月发病率。**结果**　最终拟合为ARIMA((12),1,1)模型,残差为白噪声序列,预测值与实际值的均方误差为1.742,平均绝对误差为1.159,平均绝对百分误差为0.092,2015年深圳市病毒性肝炎发病率延续了自2011年以来的逐年上升的趋势。**结论**　ARIMA模型对病毒性肝炎的时间序列变动趋势的拟合效果较好,并对未来的发病率进行预测,可为病毒性肝炎防治提供科学依据。2015年预测结果提示病毒性肝炎的发病有上升的趋势,需要进一步调整相应防控策略。

关键词:病毒性肝炎;时间序列;ARIMA模型;预测

中图分类号:R512.6;195.1　**文献标识码:**A　**文章编号:**1009-9727(2015)05-558-04

Application of ARIMA model in the prediction of viral hepatitis in Shenzhen city

ZHENG Hui-min, XUE Yun-lian, HUANG Yan-fei, DAI Chuan-wen, JIANG Shi-qiang

Nanshan District Center for Disease Control and Prevention, Shenzhen 518052, Guangdong, P.R.China

Corresponding author: ZHENG Hui-min, E-mail: vian2006@163.com

Abstract: Objective　To explore the feasibility of integrated autoregressive moving average (ARIMA) model in the prediction of the prevalence of viral hepatitis and to provide scientific basis for the control and prevention of viral hepatitis. **Methods**　The ARIMA model was developed based on monthly incidence of viral hepatitis in Shenzhen during 2004-2013 with SAS 9.2 software and it was used to predict the monthly incidence of viral hepatitis in 2014. The mean square error, mean absolute error and mean absolute percentage error of the predicted value and the actual value were measured to evaluatthe model effect then a fitting model was chosen to predict the monthly incidence of viral hepatitis in 2015. **Results**　The ARIMA model ((12),1,1) was established finally and the residual sequence was a white noise sequence. Mean square error of the predicted value and the actual value was 1.742, mean absolute error was 1.159 and mean absolute percentage error was 0.092. The incidence of viral hepatitis maintained the upward trend since 2011. **Conclusions**　The change of time series of the prevalence of viral hepatitis can be simulated with ARIMA model, which can provide evidence for the prediction of viral hepatitis. The predicted values in 2015 suggested that the prevention and control strategies on viral hepatitis should be explored further.

Key words: Viral hepatitis; Time series; ARIMA model; Forecast

病毒性肝炎是由多种肝炎病毒(目前被公认的有甲、乙、丙、丁、戊五种肝炎病毒)引起的以肝脏病变为主的一种传染病。据相关文献报道,病毒性肝炎是我国发病率居首位的传染性疾病,已成为影响我国居民健康的重要公共卫生问题,也是目前传染病防控的重点与研究的热点[1]。病毒性肝炎是深圳市发病数高,疾病负担较重的,有代表性的法定传染病[2]。深圳市的传染病年报数据显示,2005~2014年病毒性肝炎的发病数均位于深圳市法定传染病发病数的前3位。为了探讨单纯求和自回归滑动平均模型(ARIMA模型)在预测病毒性肝炎发病率方面的可行性,本文通过对深圳市2004~2014年的病毒性肝炎的月发病率进行ARIMA模型的建模拟合,预测2015年病毒性肝炎的月发病率,为制定预防控制措施提供科学依据。

1　材料与方法

1.1　*资料来源*　病毒性肝炎的月发病数据来源于2004~2014年深圳市法定传染病年报。人口资料源自深圳市统计年鉴。

1.2　*建模原理*　美国学者Box和英国统计学者Jenkins于1976年提出了博克斯-詹金斯法,简称B-J法或ARIMA模型法,它是用变量自身的滞后项以及随机误差项来解释该变量,是时间序列预测中的一种常

基金项目: 2014年度深圳市卫生计生系统科研项目(No.201402140)

作者简介:郑慧敏(1983~),女,本科,主治医师,研究方向:传染病防制。

***通讯作者:**郑慧敏,E-mail:vian2006@163.com

用而有效的方法。包括了移动平均模型(MA(q))、自回归模型(AR(p))、自回归—移动平均模型(ARMA(p,q))、求和自回归移动平均模型(ARIMA(p,d,q))、复合季节模型SARIMA(ARIMA(p,d,q)×(P,D,Q)S)及季节求和自回归—移动平均模型(ARIMA(P,D,Q)S)。其中P、D、Q分别表示季节自回归、季节差分和季节移动平均阶数,p、d、q分别表示时间序列的自回归、差分和移动平均阶数。ARIMA模型法的基本思想是将预测对象随时间推移而形成的数据视为一组依赖于时间t的随机变量,这组随机变量所具有的依存关系或自相关性表征了预测对象发展的延续性,一旦这种自相关性被识别,就可以从时间序列的过去值及现在值预测未来的值。

1.3 *建模步骤* ARIMA时间序列预测方法的建模过程有以下5个关键步骤:(1)样本的平稳化预处理:序列需为平稳化非随机性序列。(2)模型识别:包括动态数据是否为平稳序列,是否有周期性变化,是否需要差分。(3)参数估计:对识别阶段提供的粗模型进行参数估计并假设检验。(4)模型诊断:完成模型的识别、定阶和参数估计后,要对模型进行诊断以判别该模型是否能恰当地描述时间序列,对模型的有效性以及优劣做出评价。(5)预测:模型选定后,即可对将来某个时期的数值作出预测。

1.4 *统计学方法* 应用Excel软件建立数据库,SPSS 12.0和SAS 9.2实现ARIMA模型的构建和预测统计。

2 结果

2.1 *病毒性肝炎发病趋势分析* 2004~2013年深圳市的病毒性肝炎月发病率波动在2.42/10万~16.07/10万。月发病率时间序列具有明显的长期趋势和周期性,见图1。

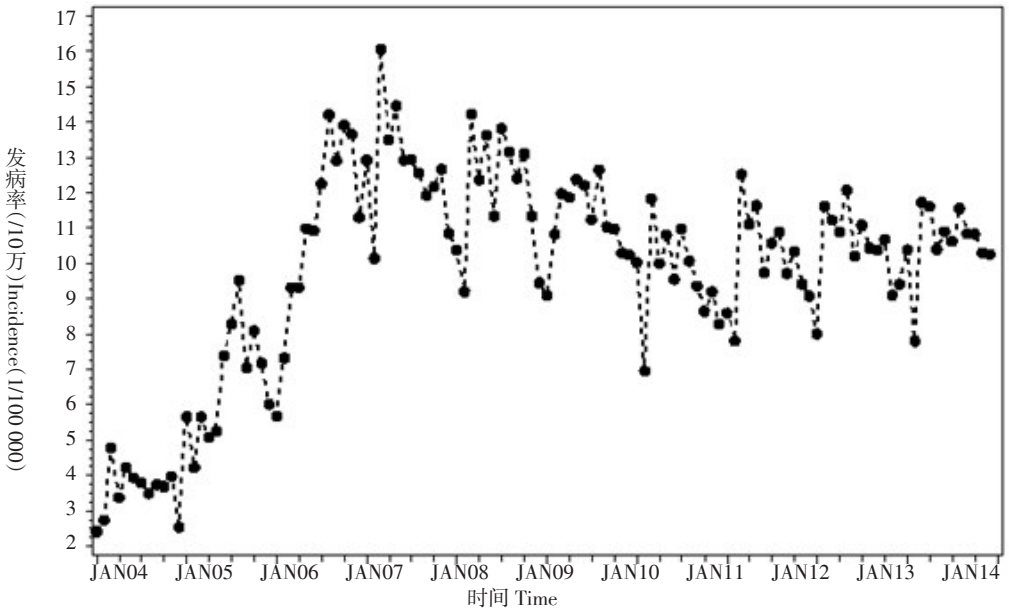

图1 2004~2013年深圳市病毒性肝炎发病率的时间序列分布

Fig. 1 Time series distribution of the incidence of viral hepatitis in Shenzhen city from 2004 to 2013

2.2 *建立模型*

2.2.1 *数据预处理* 由原始序列可知,2004~2013年病毒性肝炎的发病有长期趋势和周期性的存在,需要进行序列平稳化和非随机化的数据预处理。原始序列经过1次差分后,可见数据的线性趋势和周期性已消失,对1次差分后数据进行单位根检验,$P<0.01$,提示数据平稳,见图2。对1次差分后序列采用χ^2进行白噪声检验,P值小于显著性水平$\alpha=0.05$,所以拒绝原假设(序列为纯随机序列),接受备择假设,即序列为非随机性序列。见表1。

2.2.2 *模型识别* 对模型进行定阶,选择使AIC(Akaike Information Criterion)和SBC(Schwarz Bayesian Criterion)值相对最小,同时参数估计有统计学意义,残差为白噪声的模型。经过模型的筛选、拟和,最终选择疏系数模型ARIMA((12),1,1)进行预测。

表1 1次差分后序列的白噪声检验结果

Table 1 Autocorrelation check for white noise after the first order difference

延迟 To Lag	卡方检验 Chi-Square test	
	χ^2	P值(P value)
延迟6期 6 period lag	48.16	<0.01
延迟12期 12 period lag	68.19	<0.01

2.2.3 *参数估计和模型诊断* 用SPSS估计ARIMA((12),1,1)的参数,对模型残差序列进行白噪声检验,各阶残差序列自相关系数、偏自相关系数均落在

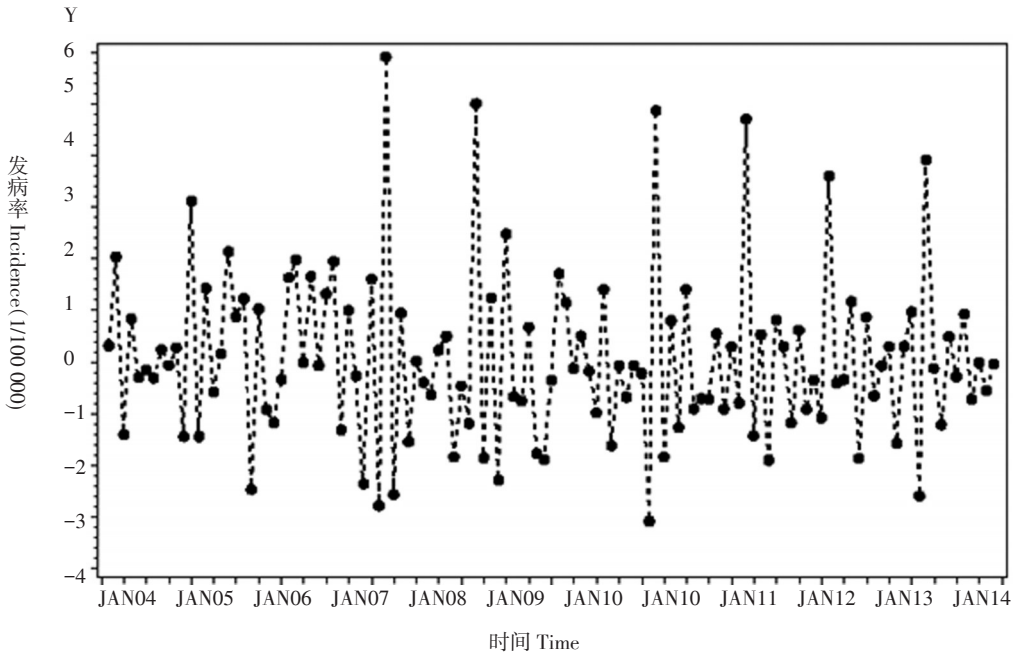

图2　2004~2013年深圳市病毒性肝炎一阶差分后时序列分布

Fig. 2　Time Series distyibution of the incidence of vural hepatits after the first order differenve in Shenzhen city from 2004 to 2013

随机区间内,残差序列为白噪声序列,说明建立的模型是合理的,可以用于预测分析。见表2。

表2　模型系数及其相关的统计量

Table 2　Coefficient of the model and related statistics

模型系数 Coefficient	估计值 Estimated value	标准差 Standard error	*t*	*P*
MU	0.07211	0.08560	0.84	0.4013
MA1,1	0.56314	0.07688	7.33	<0. 01
AR1,1	0.43641	0.08780	5.00	<0. 01

2.2.4　预测效果　根据建立的预测模型,对2014年深圳市病毒性肝炎的月发病率进行预测,预测值与实际值的MSE(均方误差)=1.742,MAE(平均绝对误差)=1.159,MAPE(平均绝对百分误差)=0.092。除个别月份外,该模型拟和数据的趋势变化和原始数据基本一致,原始数据落在拟和值的95%可信区间内。见图3。

2.2.5　2015年病毒性肝炎的月发病率预测　利用构建的ARIMA模型对深圳市2015年病毒性肝炎发病率进行预测,结果显示2015年深圳市病毒性肝炎发病率延续了自2011年以来的逐年上升的趋势,月发病率的波动形势与往年相近。随着预测时间的延长,预测值的95%可信区间变宽,预测的精度降低。这符合时间序列分析短期预测较准确,而长期预测随着未知信息的增多,估计精度越差的特点。见表3。

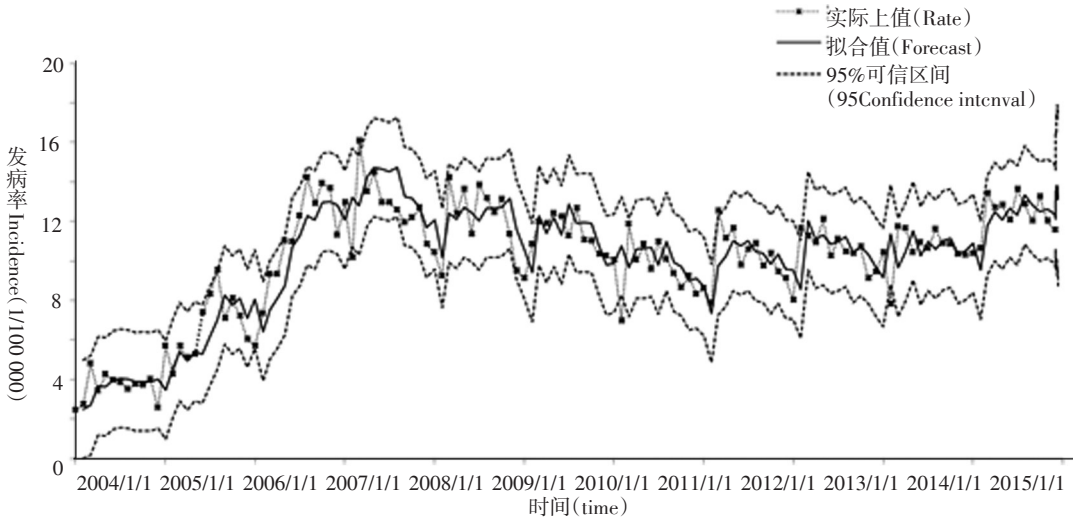

图3　2004~2014年深圳市病毒性肝炎发病率ARIMA模型拟合图

Fig. 3　ARIMA model fitting chart of incidence of viral hepatitis in Shenzhen from 2003 to 2014

表3 2015年深圳市病毒性肝炎月发病率预测

Table 3 Prediction of monthly incidence of viral hepatitis in Shenzhen in 2015

月份 Month	发病率(/100 000) Incidence(/100 000)	95%可信区间 95% Confidence interval	
		下限L95	上限U95
1月 Jan.	12.09	9.60	14.60
2月 Feb.	12.26	9.53	14.98
3月 Mar.	13.55	10.61	16.48
4月 Apr.	13.28	10.14	16.40
5月 May.	13.36	10.05	16.68
6月 June.	13.07	9.58	16.56
7月 July.	13.81	10.16	17.47
8月 Aug.	13.52	9.70	17.33
9月 Sept.	13.19	9.22	17.15
10月 Oct.	13.79	9.68	17.90
11月 Nov.	13.28	9.03	17.53
12月 Dec.	13.12	8.73	17.51

3 讨论

深圳市于2003年开始采用线性趋势移动平均法以及带趋势调整的指数平滑法开展传染病发病率预测工作,2004~2013年预测的病毒性肝炎的发病率与实际值相差甚远,平均绝对预测误差均在30%~40%之间。因此,针对当前传染病的发病情况,建立新的预测模型开展科学预测研究迫在眉睫。提高预测准确性,有效预测传染病发病率,及早发现疾病的发展趋势,为深入开展疾病的预警奠定了基础,为制定深圳市传染病防制策略提供科学依据,对传染病防制科学规划和人群防治有着重要的指导意义。

时间序列是一种通过分析预测对象的历史数据随时间发展的变化规律,建立数学模型外推的预测方法。国际预测协会主席阿姆斯特朗研究分析了57家公司的预测案例和12份公开发表的实证研究报告后发现,时间序列模型的预测效果明显优于被许多专家推崇的回归模型预测方法的效果。时间序列预测法只需要收集序列本身的历史数据,而回归模型预测法在资料的收集、汇总和分析方面有较高的要求,需要收集各种影响因素的变量值。目前我国主要的传染病监测数据为法定传染病报告数据,其他影响传染病发生、发展的各种自然、社会因素的监测数据资料并不充分。时间序列预测法符合目前我国传染病监测数据的实情,在资料收集上的成本很低,克服了回归分析法中难以掌握预测对象的影响因素和数据资料的难题,有更广泛的应用前景[3-5]。

ARIMA模型是时间序列建模中重要且预测精度较高的模型,对样本概率分布和容量没有严格要求,适用于预测难以判断变量的典型特征的数据[6-7]。本文利用了病毒性肝炎的月发病时间序列拟合了ARIMA((12),1,1)模型。分析结果显示该模型的预测精度较高,能很好地拟合原始发病序列的趋势性和周期性,可以用于病毒性肝炎发病趋势的分析和预测。

2004~2014年病毒性肝炎发病率的原始时间序列图显示:病毒性肝炎发病率于2004~2007年迅速攀升,2007~2010年逐年下降,自2011年又有上升的趋势。ARIMA模型预测2015年深圳市病毒性肝炎发病率仍将延续自2011年以来的逐年上升的趋势,因此当前病毒性肝炎的防控工作将面临严峻的压力,需调整当前的防治措施。

病毒性肝炎月发病率时间序列同时呈现明显周期性特征,每年3~10月维持在当年的一个较高水平,11月至次年1月呈下降趋势,次年2月发病率又迅速反弹。出现这种周期性波动的原因可能是受混杂因素如病人就诊量、疫情报告的质量的变化规律等影响,或者影响病毒性肝炎流行的因素与季节变化有关,具体原因尚有待于进一步分析和研究。

参考文献

[1] 朱宗元,于青.ARIMA模型在我国病毒性肝炎发病率预测中的应用[J].中国卫生统计,2011,28(1):65-67.

[2] 牟瑾,谢旭,李媛,等.将ARIMA模型应用于深圳市1980~2007年重点法定传染病预测分析[J].预防医学论坛,2009 15(11),1051-1055.

[3] 吴莹,刘文东,梁祁,等.江苏省乙型肝炎流行趋势的时间序列分析及预测[J].江苏预防医,2010,21(6):15-17.

[4] 徐国祥,胡清友.统计预测和决策[M].上海:上海财经大学出版社.1998:150-177.

[5] 严薇荣.传染病预警指标体系及三种预测模型的研究[D].武汉:华中科技大学,2008:83-84.

[6] 梁会营,李雪莲,郭军巧,等.3种模型在肾综合征出血热发病率拟合预测中的比较研究[J].中国医科大学学报,2008,37(6):843-846.

[7] 李燕婷,张宏伟,任宏,等.上海市流感样病例发病趋势的时间序列分析和预测模型研究[J].中华预防医学杂志,2007,41(6):496-498.

收稿日期:2015-02-10 **编辑**:刘雪梅