技术与方法 文章编号:0253-3626(2009)08-1070-03

# 重庆市法定报告传染病预测与监测的 ARIMA 模型

叶孟良[1],张多西[2],王润华[1]

(1.重庆医科大学公共卫生学院卫生统计与信息管理教研室,重庆 400016;2.重庆市荣昌县人民医院,重庆 402460)

【摘 要】目的:通过对 2003 年 1 月～2008 年 12 月重庆市法定报告传染病逐月发病率数据的分析,研究其变化规律,建立预测与监测的 ARIMA 时间序列模型。方法:用 Box-Ljung 统计量评价 ARIMA 模型的拟合度,用平均预测相对误差作为预测效果的评价指标。结果:重庆市法定报告传染病发病以年为周期,1 年中 4～6 月为高发月,尤其是 5 月和 6 月最为严重。ARIMA(0,1,0)(0,1,1)12 模型是重庆市法定报告传染病拟合的最佳模型,其拟合残差的方差为 12.23,外推预测的平均相对误差为 8.3%。结论:对传染病发病率历史数据进行时间序列分析是用于传染病监测的一个重要的内容。本研究所建立的 ARIMA 模型适用于重庆市传染病发病率预测与监测。

【关键词】时间序列; ARIMA 模型;预测

【中国图书分类法分类号】R181.8 【文献标识码】A 【收稿日期】2009-02-23

# Using ARIMA model to surveillance and forecast the incidence rate of notifiable infectious diseases in Chongqing

*YE Meng-liang ,et al*

(*Department of Health Statistics , College of Public Health , Chongqing Medical University*)

【Abstract】*Objective*: To develop the model for forecasting and surveilling the spread of notifiable infectious diseases in Chongqing. *Methods*: Statistics of Box-Ljung was used to evaluate the degree of fitness of ARIMA model, and the average relative errors of predict were used as indexes to evaluate the predict effect. *Results*: The changes of incidence rate of the notifiable infectious diseases in Chongqing presented a yearly periodicity, and showed that the incidence rate from March to August exceeded the monthly average of it. The residuals sum of square of the ARIMA model was 12.23 for incidence rate of the notifiable infectious disease from 2003 to 2007, and the mean relative error of the model was 8.3%.*Conclusion*: Time series methods applied to historical reporting data of infectious disease are an important tool for infectious disease surveillance. The ARIMA model is suitable to forecast the incidence rate of notifiable infectious diseases in Chongqing. Our approach potentially has a high practical value in forecasting and surveilling the notifiable infectious diseases.

【Key words】Time series; Model of autoregressive integrated moving average; Prediction

时间序列是一个依时间顺序组成的观察数据集合,应用数理统计方法加以处理,以预测未来事物的发展。移动平均法、指数平滑法是早期时间序列分析的主流方法。20 世纪 60 年代,美国学者 Box 和英国统计学者 Jenkins 提出了一整套关于时间序列分析、预测和控制的方法,被称为 Box-Jenkins 建模方法。求和自回归移动平均模型(Autoregressive integrated moving average, ARIMA)是其中重要而基本的模型之一,被大量用于时间序列资料的分析[1]。针对有季节变动规律的单因素时间序列(如传染病发病率),在本研究中,我们尝试采用混合有季节性和非季节性成分的 ARIMA 乘积模型 ARIMA(p,d,q)(P,D,Q)s,对重庆市法定报告传染病逐月发病率序列进行分析、预测和监测,探讨重庆市传染病的早期预警模型。

## 1 资料与方法

### 1.1 资料

以重庆市法定报告传染病逐月总发病率为建立模型和检验模型的原始数据。数据源于重庆市卫生局发布的 2003 年 1 月～2008 年 12 月重庆市法定报告传染病数逐月报告

作者介绍:叶孟良(1979- ),女,讲师,硕士,
研究方向:多元统计方法及其社会学应用。
通讯作者:王润华,男,教授,E-mail:wzhzyp@163.com。

资料和重庆市统计局报告的重庆市当年人口数。2003 年 1 月～2007 年 12月重庆市法定报告传染逐月病发病率的数据用于建立模型,2008 年 1 月～12 月相应数据用于模型检验。

1.2 方法

1.2.1 时间序列模型 根据重庆市 2003 年～2008 年法定报告传染病发病率序列图,判定该序列的变化规律。利用自相关分析初步确定该序列适宜的模型和模型参数。建立求和自回归移动平均模型,分析发病率的长期趋势、季节性、周期性变化和随机干扰,并对法定报告传染病的发病率进行预测。平稳序列的 ARMA(p,q)模型为 $Y_1=\phi Y_{t-1}+\phi_2 Y_{t-2}+\cdots\phi_p-Y_{t-p}+e_t-\theta_1 e_{t-1}-\theta_2 e_{t-2}\cdots-\theta_q e_{t-p}$ 其中,$\phi_1,\phi_2\cdots\phi_p;\theta_1,\theta_2\cdots\theta_q$ 为待估计的参数,p,q 称为自回归和移动平均的阶数。对于非平稳序列,可以通过对原序列进行非季节或季节差分,使得序列平稳化,模型为 ARIMA(p,d,q)(P,D,Q)s,其形式为: $\phi_p(B)\phi_p(B^s)\nabla^d\nabla^D x_t=\theta_q(B)\Theta_Q(B^s)e_t$ 其中 p,d 和 q 分别表示非季节自回归阶数、差分次数和滑动平均的阶数;P,D 和 Q 分别表示季节自回归阶数、季节差分次数和季节滑动平均阶数[1]。

1.2.2 模型选择和参数估计 利用非线性最小二乘法估计模型参数。ARIMA 模型的诊断从以下几方面进行:①模型参数是否有统计学意义;②备选模型的拟合优度比较;③参数独立性;④残差检验[2]。利用 2008 年 1 月～2008 年 12 月数据对模型的外推效果进行检验,利用平均相对误差评价模型的外推预测效果。所有计算由SPSS13.0 软件实现。

1.2.3 病人收容量预测和监测 用筛选的最优 ARIMA 模型,对重庆市传染病发病率进行预测,通过模型拟合值的 95%置信区间绘制发病率监测图。

1.2.4 与指数平滑法的比较 指数平滑法是对近期序列值赋予较大的权重,来估计序列的平均水平、趋势性和周期性。用筛选的最优 ARIMA 模的预测效果与指数平滑法的预测效果进行比较。

## 2 结 果

### 2.1 重庆市法定报告传染病发病率变化规律

重庆市法定报告传染病发病率的序列图(图 1)显示 2003~2008 年逐月发病率序列有明显的季节性,且有一个微弱增加趋势,并存在周期性,周期约为 12 个月,其中 2003 年 12 月发病率最低为 8.60/10 万,2008 年 5 月发病率最高为 58.21/10 万。自相关图和偏相关图(图 2)显示发病率序列存在自回归和滑动平均过程。

表 1 2003~2008 年重庆市法定报告传染病季节调整指数

| 时间(月) | 季节调整指数(%) |
|---|---|
| 1 | 96.8 |
| 2 | 83.1 |
| 3 | 106.1 |
| 4 | 113.4 |
| 5 | 125.8 |
| 6 | 120.4 |
| 7 | 109.3 |
| 8 | 105.2 |
| 9 | 97.8 |
| 10 | 96.1 |
| 11 | 83.3 |
| 12 | 62.5 |

季节调整指数[3]分析也显示重庆市法定传染病发病率序列存在季节效应,其中 3~8 月发病率超过了重庆市这 6 年的月平均发病率,其中 5 月平均发病率最大。见表 1。

### 2.2 ARIMA 预测模型参数估计

根据重庆市 2003 年 1 月~2007 年 12 月法定报告传染病发病率时间序列,建立模型见(表 2),所列模型参数均有统计学意义,且残差通过了白噪声检验。通过模型的筛选,得出 ARIMA(0,1,0)(0,1,1)12 为最优模型。拟合效果如图 1 所示,外推预测结果见表 3,平均相对误差为 8.3%。

### 2.3 重庆市法定报告传染病发病率的监测

利用 2003 年~2007 年重庆市法定报告传染病发病率数据建立 ARIMA 模型,并预测 2008 年逐月发病率的值,以模型拟合值的 95%置信限绘制控制线,用于重庆市法定报告传染病疫情的监测。见图 1。

### 2.4 比较

有季节规律的时间序列用线性或非线性回归模型进行预测显然是不合适的。这里尝试用指数平滑的方法进行建模[4],得模型的常规参数 $\alpha=0.7$,趋势参数 $\gamma=0.0,\delta=1.0$。指数平滑发预测的平均相对误差为 16.4%,而 ARIMA(0,1,0)(0,1,1)12 平均相对误差为 8.3%,显然对于重庆市法定报告传染病的 ARIMA 模型的预测效果较佳。

表 2 重庆市法定报告传染病发病率备选模型拟合优度统计量

| 模型 | SE | Log-likelihood | AIC | BIC |
|---|---|---|---|---|
| ARIMA(0,1,0)(0,1,1)12 | 3.523 | -127.454 | 258.909 | 262.609 |
| ARIMA(1,0,0)(1,1,0)12 | 3.550 | -129.232 | 264.464 | 270.078 |
| ARIMA(1,1,1)(1,0,0)12 | 3.778 | -166.689 | 341.378 | 349.688 |

表 3 重庆市法定报告传染病发病率(1/10 万)模型外推预测结果

| 时间(月) | 实际发病率 | 预测发病率 | 相对误差 | 平均相对误差 |
|---|---|---|---|---|
| 1 | 37.68 | 39.31 | 0.017 | 0.083 |
| 2 | 30.02 | 31.91 | 0.063 | |
| 3 | 42.72 | 39.39 | 0.078 | |
| 4 | 45.05 | 43.03 | 0.045 | |
| 5 | 58.21 | 46.44 | 0.202 | |
| 6 | 52.90 | 44.55 | 0.158 | |
| 7 | 43.87 | 41.11 | 0.063 | |
| 8 | 37.42 | 41.09 | 0.098 | |
| 9 | 35.69 | 39.09 | 0.095 | |
| 10 | 36.52 | 37.10 | 0.016 | |
| 11 | 35.20 | 32.91 | 0.065 | |
| 12 | 28.32 | 25.64 | 0.095 | |

## 3 讨 论

ARIMA 乘积模型是针对有季节性变动的时间序列提出的建模方法。它对每一季节周期中相同时间点的序列值进行分析,提取季节趋势,并针对每个季节周期内部序列值的变化提取非季节性成分,确立最优模型。该模型一般能够较好地描述时序的变动情况,特别适合于时序规律不明显,或有明显季节性和周期性的情况[5]。时间序列分析中指数平滑法是用无限大的宽窗,各历史值的权重随时间的推移呈指数衰减来估计序列的平均水平、趋势性和周期性。但通过比较,可以看出 ARIMA 乘积模型的预测平均相对误差为 8.3%,而指数平滑法为 16.4%,ARIMA 乘积模型的预测效果优于指数平滑法。ARIMA$(0,1,0)(0,1,1)_{12}$ 模型对 2008 年的预测值与实际发病率吻合得较好,能够起到预测预报的作用。由于预测所用历史资料源于法定报告传染病的报告系统,其完整性和可信度各地差异很大,因此,本文所建立的预测模型也存在着一定的地域局限性。由于时间序列预测是按照一定历史时期的规律进行,其前提是这些规律在一定时期内保持相对稳定,预测时间不宜太长。因此,可通过逐年数据的补充,以及不断改进模型的估计方法,对模型进行修正以提高预测模型精度。

从以上分析结果可知,重庆法定报告传染病发病率序列有微弱上升的趋势,这与 2004 年开始实行医疗机构网络报告,提高了报告率有关[6]。重庆市法定报告传染病发病率还表现为明显的季节性,5~6 月出现高峰,同时存在周期性,周期约为 12 个月,这可能与报告的发病数顺位中前几位病原体生长繁殖存在周期性、季节性有关。

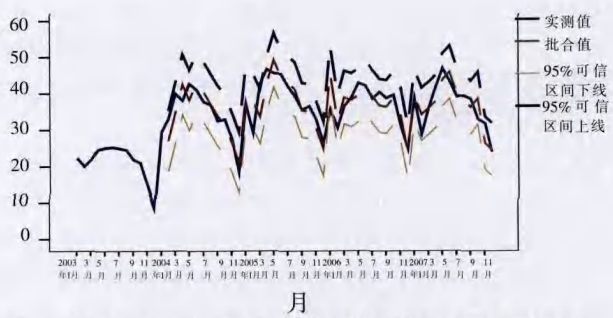



图 1 2003~2007 年重庆市法定报告传染病发病率序列图

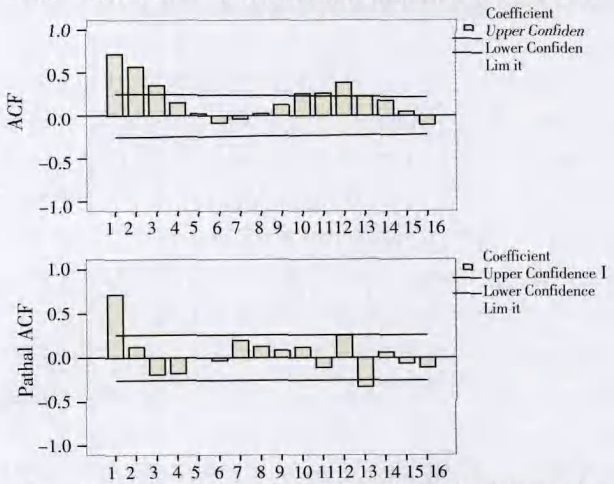



图 2 2003~2007 年重庆市法定报告传染病发病率 ACF 图与 PACF 图

摸清传染病流行规律,建立预测模型可以为有关部门采取防治措施提供依据。传染病防治部门应根据传染病流行的季节变化规律、预测数据有针对性开展防治工作;在疫情监测工作中,根据预测数据的可信限,可以判断实际发病率是否在正常范围波动,应警惕传染病暴发或流行的可能,从而全面提高传染病综合防治能力和水平,有效降低传染病发病率。

## 参 考 文 献

[1] 宇传华,曹 阳,方 亚,等.宇传华.SPSS 与统计分析[M].第一版.北京:电子工业出版社,2007.578-600.

[2] 张文彤.SPSS 统计分析教程[M].第 1 版.北京:北京希望电子出版社,2002.250-289.

[3] 冯刘栋.试用 3 次指数平滑法预测传染病发病率[J].数理医药学杂志,2000,13(2):145-146.

[4] 刘金莲,钟晓妮.ARIMA 季节乘积模型在预测医院月住院量的应用[J].重庆医科大学学报,2008,33(4):475-477.

[5] 冯 丹,韩晓娜,赵文娟,等.中国内地法定报告传染病预测和监测的 ARIMA 模型[J].疾病控制杂志,2007,11(2):140-143.

[6] 王小英.潮州市 1994~2005 年甲乙类传染病疫情分析[J].中国热带医学,2006,6(11):1994-1995.

(责任编辑:冉明会)