·现场调查·

不同时间序列分析法在洞庭湖区血吸虫病发病预测中的比较

赛晓勇 张治英 徐德忠 闫永平 李良寿 蔡凯平 李岳生 周晓农

【摘要】 目的 通过比较时间序列分析中指数平滑法、移动平均法、自回归分析及自回归综合移动平均法(ARIMA)在洞庭湖区退田还湖濠口试点 1990～2002 年血吸虫病患病率预测中的优劣。方法 用时间序列分析各方法建模预测,比较各方法 1994～2002 年预测值的误差平方和,确定最佳预测方法。结果 指数平滑法、移动平均法、自相关分析及 ARIMA 法中 1994～2002 年预测值的误差平方和依次为39.40、39.86、26.63、22.54。结论 濠口试点 1990～2002 年患病率预测中,时间序列分析诸方法中 ARIMA 模型预测效果较好。

【关键词】 血吸虫病;时间序列分析;统计预测

Application of "time series analysis" in the prediction of schistosomiasis prevalence in areas of "breaking dikes or opening sluice for waterstore" in Dongting Lake areas, China

SAI Xiao-yong, ZHANG Zhi-ying, XU De-zhong, YAN Yong-ping, LI Liang-shou, CAI Kai-ping, LI Yue-sheng, ZHOU Xiao-nong. *Department of Epidemiology, Faculty of Preventive Medicine, Fourth Military Medical University, Xi'an* 710033, China

Corresponding author: *XU De-zhong*, Email: xudezh@fmmu.edu.cn

【**Abstract**】 **Objective** To provide the fittest model for forecasting schistosomiasis prevalence in Haokou village of "breaking dikes or opening sluice for waterstore" in Dongting Lake areas by comparing the results of Moving Average, Exponential Smoothing, Autoregressive Model and Autoregressive integrated moving average model (ARIMA model) from 1990 to 2002. **Methods** Error sum of square of four statistical methods was compared and the fittest model was chosen. **Results** Error sum of square of predicted schistosomiasis prevalence rates in Haokou village from 1994 to 2002 were 39.40, 39.86, 26.63, 22.54 respectively. **Conclusion** ARIMA model seemed to be the fittest one in the prediction of schistosomiasis prevalence in Haokou village of "breaking dikes or opening sluice for waterstore" in Dongting Lake from 1990 to 2002.

【**Key words**】 Schistosomiasis; Time series analysis; Statistical prediction

2003 年 8 月 25 日我国卫生部宣布试行《血吸虫病重大疫情应急处理预案》,主要因为南方各省入夏后血吸虫病患者增多,疫情有了新的变化。血吸虫病发病影响因素,如气侯、洪水、钉螺等的变化使得血吸虫病发病随时间出现了一定程度的周期性变化[1-3],我们应用时间序列分析对国家"十五"课题湖南洞庭湖区退田还湖试点1990～2002 年的血吸虫病疫情资料进行分析,以阐明其变化规律,达到快速有效的预测。

基金项目:国家"十五"科技攻关课题资助项目(2001BA705B08)

作者单位:710033 西安,第四军医大学预防医学系流行病学教研室(赛晓勇、张治英、徐德忠、闫永平、李良寿);湖南省血吸虫病防治所(蔡凯平、李岳生);中国疾病预防控制中心寄生虫病预防控制所(周晓农)

通讯作者:徐德忠,Email:xudezh@fmmu.edu.cn

资料与方法

1. 资料来源:收集洞庭湖区退田还湖澧县的濠口试点(单退点,即退人不退田,洪水期人转移、洪水过后返回种田)1990～2002 年连续粪检阳性率的病情资料。选择历年的粪检阳性率(资料全),由随机抽样调查而来。濠口试点1990～2002 年常住人口 1176 人,面积为2 970 000 m^2,为湖南省血吸虫病重灾区监测试点之一。全部资料由湖南省血吸虫病防治(血防)所及小渡口血防站提供。

2. 方法:时间序列分析是根据被预测变量自身的变化规律来建立模型,然后利用这个模型来预测该变量未来的变化。时间序列分析包括指数平滑法、移动平均法、自回归分析及自回归综合移动平均

法(ARIMA 法)。评价主要是通过比较各方法的拟合优度、误差平方和实现。分析由 SPSS 11.0软件完成。

3. 建模、预测:

(1)移动平均法:是利用一组观察值的均值作为下一期的预测值,设时间序列为 $x_1, x_2, x_3, \cdots$,可以表示为 $F_{t+1} = \frac{1}{N}\sum_{t-N+1}^{t} x_i$,式中 x_t 为最新观察值;F_{t+1}为下一期的预测值,N 为一组观察值的个数。q 阶移动平均模型的公式为:

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

用自相关系数识别,它的自相关系数为:

$$r_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \Lambda + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \Lambda + \theta_q^2} & 1 \leqslant k \leqslant q \\ 0 & k > q \end{cases}$$

时间序列相差 k 个时期两项数据序列之间的依赖程度可用自相关系数 r_k 表示为:

$$\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y}) \Big/ \sum_{t=1}^{n}(Y_t - \overline{Y})^2$$

式中:n 是时间序列 Y_t 的数据个数;Y_{t-k}是其滞后 k 期数据形成的序列。$\overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$ 是时间序列的平均值。r_k 取值范围在正负 1 之间,$|r_k|$与 1 越接近,说明时间序列的自相关程度越高。

(2)指数平滑法:用序列过去值的加权均数来预测将来的值,并给近期的更大的权数,远期的给以较小的权数。表达式为:$\hat{z}_{t+1} = \alpha z_t + (1-\alpha)\hat{z}_t$,$\alpha$ 为平滑指数,$\hat{z}_{t+1}$为下一年预测值,z_t 为当年真实值,$\hat{z}_t$ 为当年预测值。到时期 t 时,只需知道实际数值和本期预测两个数据值就可预测下一个时间的数值。

(3)自回归分析:自回归分析主要是对时间序列求其本期与不同滞后期的一系列自相关系数和偏自相关系数以识别其特性,主要用偏自相关系数来判定模型的阶数。P 阶自回归 AR(P)模型的公式为:

$$Y_t = \varnothing_1 Y_{t-1} + \varnothing_2 Y_{t-2} + \cdots + \varnothing_p Y_{t-p} + e_t$$

它的偏自相关系数满足:

$$\varnothing_{ki} = \begin{cases} \varnothing_i & 1 \leqslant i \leqslant p \\ 0 & p+1 \leqslant i \leqslant k \end{cases}$$

偏自相关是时间序列 Y_t 在给定了 $Y_{t-1}, Y_{t-2}, \cdots, Y_{t-k+1}$的条件下,$Y_t$ 与滞后 k 期时间序列之间的条件相关。它用来度量当其他滞后 1,2,3,…,$k-1$ 期时间序列的作用已知的条件下 Y_t 与 Y_{t-k}之间的相关程度,用 Φ_{kk} 度量。$\Phi_{kk} = (r_k - \sum_{i=1}^{k-1}\Phi_{k-1}, i \times r_{k-i}) \Big/ (1 - \sum_{i=1}^{k-1}\Phi_{k-1}, i \times r_i)$。$k = 2, 3, \cdots$,式中:$\Phi_{k,i} = \Phi_{k-1,i} - \Phi_{kk} \times \Phi_{k-1,k-i}$,$i = 1, 2, \cdots, k-1$。

(4)ARIMA 模型:首先判定数据有无随机性、平稳性、季节性,然后要在预测之前实现最优拟合、建模,最后进行预测及评价。模型为 ARIMA(p,d,q),它将移动平均、自回归分析及差分结合起来。确定 3 个参数,即自回归阶数(p)、差分次数(d)、移动平均阶数(q),它首先通过差分把时间序列的季节性消除之后(达到数据平稳),然后建模,最后估计参数。对非季节数据,一般求一阶差分即可。若时间序列的季节性的变动周期为 T,时间序列 Y_t 的一阶季节差分序列$\nabla_T Y_t$ 为$\nabla_T Y_t = Y_t - Y_{t-T}$($t > T$)。自相关分析图将自相关系数和偏自相关系数绘制成图,并标出了置信区间,利用它我们可分析时间序列的随机性、平稳性和季节性。随机性是指时间序列各项之间没有相关关系的特性。判定准则:自相关系数基本上落在置信区间内。平稳性是指时间序列的统计特征不随时间推移而变化。判定准则:自相关系数 r_k 在$k > 3$ 时都落入置信区间内并逐渐趋于零。季节性是指在某一固定时间间隔上,重复出现的某种特性。判定准则:某一时间序列在 $k = 2$ 或 3 以后的自相关系数 r_k 值存在着周期性的显著不为零的值,则有季节性[4]。

4. 各方法评价:比较各方法 1994~2002 年预测值与真实值的误差平方和(SSE),较小者为优。

结 果

1. 建模、预测:

(1)用移动平均法建模预测:我们以最常用的 3 年为周期,代入$y = (1.5x_t + x_{t-1} + 0.5x_{t-2})/3$ 计算得1994~2002 年粪阳率预测值。y 为预测下一年的值,x_t、x_{t-1}、x_{t-2}分别为当年、上一年及前年的粪阳检率实测值。

(2)用指数平滑法建模预测:首先通过对数据的平稳性分析,我们选用简单指数平滑法,确定使得 SSE 最小的 α 值为0.9(SSE = 26.9675)如表 1 所示。代入$y = 0.9x_t + 0.1x_{t估}$,y:预测下一年的值,x_t:当年实测值,$x_{t估}$:当年估计值。

(3)用自相关分析建模预测:通过做自相关分析中的散点图发现濠口粪检阳性率有直线趋势($F =$

231.1, $P<0.05$)(图 1)。其自回归方程为 $y=0.953x$($R=0.959$, $P<0.05$),其中,y :预测下一年的值,x :当年实测值。

表1 指数平滑法 SSE 比较

自由度	α 值	SSE
10	0.9	26.9675
	0.8	28.9914
	0.7	31.3835
	0.6	34.0854
	0.5	36.9129

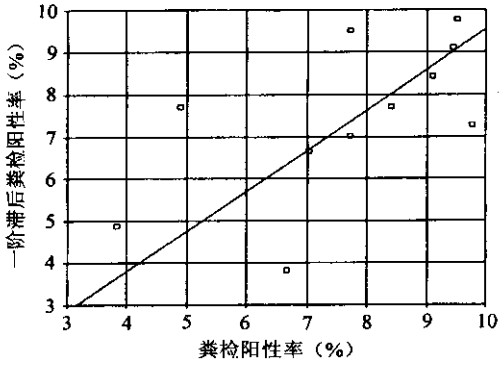

图1 粪检阳性率一阶滞后散点图

(4)用 ARIMA 模型建模预测:进行分析时,首先要确定 ARIMA 模型各阶数,确定阶数通过比较不同阶数时拟合优度及分析自相关分析图和偏自相关分析图实现。如表 2、图 2、3 显示 ARIMA(1,1,1)较好,拟合优度统计量包括标准误(standard error)、对数似然函数值(log likelihood)、Akaike 信息准则值(AIC)、Schwarz 判定准则值(SBC)[5]。

表2 不同 ARIMA 模型拟合优度的比较

统计量	ARIMA				
	2,1,2	1,1,2	1,1,1	2,1,1	2,2,2
残差数	10	10	10	10	9
标准误	1.6816	1.5735	1.6036	1.4636	1.8284
对数似然函数值	−17.2206	−17.1063	−17.4884	−16.6343	−16.3016
Akaike 信息准则值	44.4413	42.2126	40.9768	41.2686	42.6032
Schwarz 判定准则值	45.9542	43.4230	41.8846	42.4790	43.5893

注:ARIMA(1,1,1):自回归阶数(p)为 1、差分次数(d)为 1、移动平均阶数(q)为 1 时的 ARIMA 模型,余同

代入公式得到预测方程。预测方程为:$\hat{y}_t=0.6487Y_{t-1}+0.3513Y_{t-2}+0.8671e_{t-1}+0.0223$。$\hat{y}_t$:预测当年的值,Y_{t-1} 为上一年的实测值,余类推。e_{t-1} :上一年预测值的误差。

2. 四种方法预测效果比较:通过计算 1994～2002 年预测误差平方和判断预测效果优劣,结果如表 3 所示。

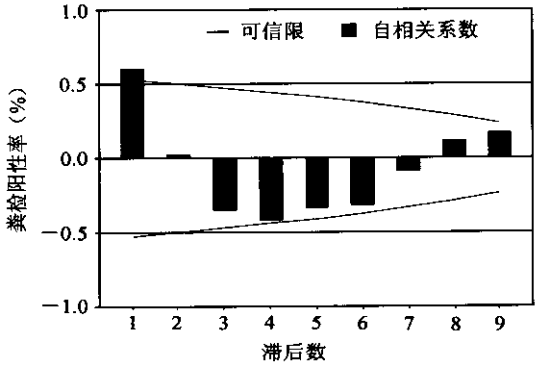

图2 自相关分析图

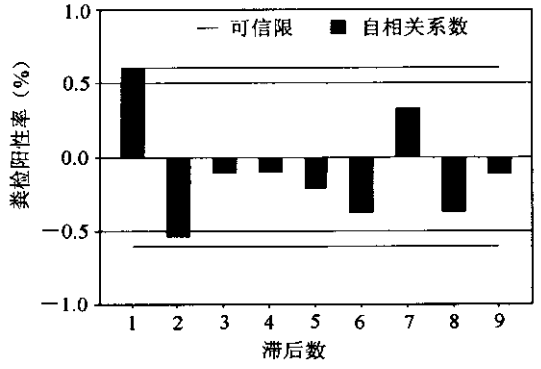

图3 偏自相关分析图

表3 1994～2002 年四种方法粪检阳性率预测效果比较

方法	SSE
指数平滑法	39.40
移动平均法	39.86
自相关分析	26.63
自回归综合移动平均法	22.54

讨 论

公共卫生事件中及时、快速地预警能够避免国民经济遭受更大的损失。在血吸虫病的防治工作中,如何准确、及时地预测血吸虫病的发病成为卫生机构决策者的难题,目前还没有有效、成熟的方法。

时间序列分析逐渐被用于医学研究领域,国外应用广泛,Diaz 等[6]用 ARIMA 模型研究温度与死亡率的关系时发现,75 岁以上组在 41℃ 以上时,每升高一度全因死亡率超过平均达 51% 以上。McCleary 等[7]曾运用时间序列分析阐述自杀与赌博的关系发现其不相关。Clancy 等[8]运用时间序列分析认为空气颗粒污染的控制可以减少呼吸道和心血管疾病的死亡率。国内已有人运用时间序列分析对恶性肿瘤、肾综合征出血热、乙型脑炎、血吸虫病等进行了初步研究[9-13],而系统地运用时间序列各

种方法对血吸虫病的发病进行预测还未见报道。利用时间序列模型不需要知道影响预测变量的相关因素,这是其他预测方法所不能比拟的,可以通过既往资料快速预测。但也应看到,正是由于未考虑影响预测变量的相关因素,它也有局限性,适合于受预测变量的相关因素影响较小的试点,即相对稳定的试点,如濠口。

时间序列分析法各有特点。移动平均法有两个优点:一是计算量少,如本例以 3 年为周期,只需连续 3 年数据即可预测;二是移动平均线能较好地反映时间序列的趋势及变化。但它有两个限制,一是必须有 N 个过去观察值,如本例必须有连续 3 年资料;二是过去观察值中每一个权数都相等,早于 $t-N+1$ 的观察值权数为零。指数平滑法需要通过反复试验确定使均方差最小的 α 值,本例确定的 α 值为 0.9,方程为 $y=0.9x_t+0.1x_{t估}$,它只需知道上一年的资料即可。自相关分析依赖于样本量,必须有一组连续变量。而 ARIMA 法将移动平均法、自相关分析及数据的平稳性考虑在了一起,通过自相关系数和偏自相关系数分析确定 q 和 p。四种方法中,理论上讲 ARIMA 法更全面,综合考虑因素多,本研究也验证了这一点,但在不同的应用条件下,还要视具体情况而定,如王谦等[14,15]通过对四川省 41 个贷款县血吸虫病流行变化的规律研究认为移动平均数法预测的结果较好,并采用移动平均数法预测得出 1999~2001 年主要血吸虫病流行区人群血吸虫病感染率将缓慢上升、患者数将逐渐增加、耕牛血吸虫病感染率和病牛数将呈反复波动趋势、有螺面积将逐年小幅下降的结论。

本课题曾对洞庭湖区退田还湖试点 1990~2002 年血吸虫病情与螺情进行分析,发现了血吸虫病及活螺密度退田还湖前后的变化趋势,但不能进行定量预测[16],本研究则为定量预测提供了有效工具。时间序列模型预测的偏差大小受数据本身特点、样本量大小等因素影响。本研究结论,在濠口试点1990~2002 年患病率预测中,ARIMA 模型预测效果较好。

参 考 文 献

1 洪青标,周晓农,孙乐平,等. 全球气候变暖对中国血吸虫病传播影响的研究——钉螺越夏致死高温与夏蛰的研究. 中国血吸虫病防治杂志,2003,15:24-26.

2 李涛,余秉圭,戴裕海,等. 长江洪水对急性血吸虫病流行的影响及防治对策研究. 中国血吸虫病防治杂志,2000,12:268-272.

3 孙乐平,周晓农,洪青标,等. 长江下游江滩地区血吸虫病再流行规律的研究——钉螺的迁入与消长. 中国血吸虫病防治杂志,2001,13:213-215.

4 张文彤,主编. SPSS11 统计分析教程(高级篇). 北京:北京希望电子出版社,2002. 284.

5 徐国祥,主编. 统计预测与决策. 上海:上海财经大学出版社,1998. 158-162.

6 Diaz J, Garcia R, Velazquez, et al. Effects of extremely hot days on people older than 65 years in Seville (Spain) from 1986 to 1997. Int J Biometeorol, 2002, 46:145-149.

7 McCleary R, Chew KS, Merrill V, et al. Does legalized gambling elevate the risk of suicide? An analysis of U. S. counties and metropolitan areas. Suicide Life Threat Behav, 2002, 32:209-210.

8 Clancy L, Goodman P, Sinclair H, et al. Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study. Lancet, 2002, 360:1210-1214.

9 吴进军,苏汝好. 四会市鼻咽癌发病率及死亡率时间序列分析与预测. 中国卫生统计,2000,17:345-348.

10 侯世方,孙长福,王毅,等. 简易季节时间序列分析法的应用. 中国卫生统计,2001,18:176-177.

11 吴进军,苏汝好. 中山市宫颈癌发病率及死亡率时间序列分析与预测. 医学信息,2000,13:569-572.

12 李来英,张小平. 起伏型时间序列分析方法在流行性出血热预测中的应用. 中国卫生统计,1997,14:64.

13 李廷杰,陈秀山,李燕芬. 应用时间序列统计方法分析广东省 1984~1993 年乙型脑炎季节性分布特性. 中华流行病学杂志,1998,19:103-106.

14 王谦,肖永富,蒋朝东. 防治过程中血吸虫病流行趋势预测方法研究. 实用寄生虫病杂志,1999,7:120-121.

15 肖永富,王谦,魏继炳. 四川省 1999~2001 年血吸虫病流行趋势预测. 实用寄生虫病杂志,2000,8:53-55.

16 赛晓勇,蔡凯平,徐德忠,等. 洞庭湖区退田还湖试点 1990/2002 血吸虫病情与螺情分析. 第四军医大学学报,2003,24:1878-1880.

(收稿日期:2003-09-29)

(本文编辑:尹廉)