·研究原著·　　文章编号:1000-2790(2003)24-2297-04

时间序列分析在洞庭湖区双退试点血吸虫病发病预测中的应用

赛晓勇[1],张治英[1],徐德忠[1],闫永平[1],蔡凯平[2],李岳生[2],周晓农[3]　([1]第四军医大学预防医学系流行病学教研室,陕西西安 710033,[2]湖南省血吸虫防治研究所,湖南 岳阳 414000,[3]中国疾病预防控制中心寄生虫病预防控制所,上海 200032)

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

SAI Xiao-Yong[1], *ZHANG Zhi-Ying*[1], *XU De-Zhong*[1], *YAN Yong-Ping*[1], *CAI Kai-Ping*[2], *LI Yue-Sheng*[2], *ZHOU Xiao-Nong*[3]

[1]Department of Epidemiology, School of Preventive Medicine, Fourth Military Medical University, Xi'an 710033, China, [2]Hunan Institute of Anti-epidemic of Schistosomiasis, Yueyang 414000, China, [3]Institute of Parasitic Diseases, Chinese Center for Disease Control & Prevention, Shanghai 200032, China

【Abstract】AIM: To provide the fittest model for prediction of schistosomiasis prevalence in Jicheng of "breaking dikes or opening sluice for waterstore" in Dongting Lake by comparing the results of moving average, exponential smoothing, autoregressive model and autoregressive integrated moving average model (ARIMA model) from 1990 to 2002, to provide reference to schistosomiasis preventive work. **METHODS**: The fittest model was chosen by comparing the goodness of fit, error sum of square and relative error of four statistical methods. **RESULTS**: The average predicted relative errors and error sum of square of ARIMA from 1993 to 2002 are the smallest. **CONCLUSION**: The fittest model in the prediction of schistosomiasis prevalence in Jicheng of "breaking dikes or opening sluice for waterstore" in Dongting Lake from 1990 to 2002 is ARIMA model.

【Keywords】 time series analysis; forecasting; schistosomiasis

【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(Autoregressive integrated moving average model, ARIMA model)在洞庭湖区退田还湖试点集成垸1990/2002年血吸虫病患病率预测中的优劣,为当地退田还湖试点的血吸虫病发病找到一个较适合的预测模型,为防治工作提供参考依据. **方法**:用时间序列分析各方法建模预测,比较各方法的拟合优度、误差平方和及预测值的相对误差,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法的1993/2002年患病率预测值年平均相对误差(%)和误差平方和ARIMA模型最小. **结论**:集成垸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】时间序列分析;预测;血吸虫病
【中图号】R181.8　　**【文献标识码】**A

0　引言

2003-08-25卫生部宣布试行《血吸虫病重大疫情应急处理预案》,主要原因是我国南方各省今年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的变化,如气侯、洪水、钉螺等变化,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年的病情资料进行分析,以期阐明其变化规律,达到快速有效预测,为国家卫生机构提供决策依据.

1　材料和方法

1.1　材料　收集洞庭湖区退田还湖澧县的集成垸试点(双退点,即退人又退田,该院1998年后完全废弃用于泄洪)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,病情资料由每年随机抽样调查而来. 集成垸试点退田还湖后滞留人口2600人,面积为2200万平方米,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防治研究所及华容县小渡口血吸虫防治站提供.

1.2　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. 包括指数平滑法、移动平均法、自回归分析及ARIMA法. 评价主要是通过比较各方法的拟合优度、误差平方和及预测值的相对误差实现. 统计分析由SPSS11.0软件完成.

1.2.1　建模、预测　① 移动平均法:利用一组观察

收稿日期:2003-08-25;　修回日期:2003-10-22
基金项目:国家"十五"科技攻关课题(2001BA705B08)
通讯作者:徐德忠. Tel. (029)3374955　Email. xudezh@fmmu.edu.cn
作者简介:赛晓勇(1974-),男(回族),河南省新乡市人. 讲师,硕士生(导师徐德忠,闫永平). Tel. (029)3374871 Ext. 11　Email. saixiaoyong@163.com

万方数据

值的均值做为下一期的预测值,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:$Y_t=e_t-\theta_1e_{t-1}-\theta_2e_{t-2}-\cdots-\theta_qe_{t-q}$,用自相关系数识别,它的自相关系数为:$r_k=\begin{cases}\dfrac{-\theta_k+\theta_1\theta_{k+1}+\cdots\theta_{q-k}\theta_q}{1+\theta_1^2+\theta_2^2+\cdots+\theta_q^2} & 1\leqslant k\leqslant q\\ 0 & k>q\end{cases}$. 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 $\sum_{t=k+1}^{n}(Y_t-\bar{Y})(Y_{t-k}-\bar{Y})/\sum_{t=1}^{n}(Y_t-\bar{Y})^2$. 式中:$n$ 是时间序列 Y_t 的数据的个数;Y_{t-k}是其滞后 k 期数据形成的序列. $\bar{Y}=\frac{1}{n}\sum_{t=1}^{n}Y_t$,是时间序列的平均值. r_k 取值范围为 -1 到 1,$|r_k|$ 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,远期的序列值给以较小的权数. 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值两个数据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:$Y_t=\phi_1Y_{t-1}+\phi_2Y_{t-2}+\cdots+\cdots\phi_pY_{t-p}+e_t$,它的偏自相关系数满足:$\phi_{kk}=\begin{cases}\phi i & 1\leqslant i\leqslant p\\ 0 & p+1\leqslant i\leqslant k\end{cases}$、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. 它用来度量在滞后 1,2,3,$\cdots$,$k-1$ 期时间序列作用已知的条件下 Y_t 与 Y_{t-k}之间的相关程度,用 Φ_{kk}度量. $\Phi_{kk}=(r_k-\sum_{i=1}^{k-1}\phi_{k-1},\ i\times r_{k-1})/(1-\sum_{i=1}^{k-1}\phi_{k-1},\ i\times r_i)$. $k=2,3,\cdots$式中:$\Phi_{k,i}=\Phi_{k-1,i}-\Phi_{kk}\times\Phi_{k-1,k-i}$,$i=1,2,\cdots,k-1$. r_k 和 r_i 表示 k 期和 i 期的自相关系数. 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 ∇_TY_t 为 $\nabla TY_t=Y_t-Y_{t-T}(t>T)$. 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].

1.2.2 *各方法评价* 比较各方法 1993/2002 年预测值的年平均相对误差和年预测值与真实值的误差平方和,较小者为优.

2 结果

2.1 建模、预测

2.1.1 *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 2002 年粪检阳性率预测值为 26.66%. y 为预测下一年的值,x_t、x_{t-1}、x_{t-2}分别为当年、上一年及前年的粪检阳性率实测值.

2.1.2 *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 = 192.88),如 Tab 1 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. 计算得2002 年粪阳率预测值为 30.53%.

表 1 指数平滑法不同 α 值时最小误差比较

Tab 1 Comparison of SSE in different α

DFE	α value	SSE
10	0.9	192.88
	0.8	197.67
	0.7	209.93
	0.6	231.57
	0.5	265.18

SEE: Sum of square error; DFE: Degree of freedom error.

2.1.3 *用自相关分析建模预测* 做自相关分析中的散点图发现集成粪检阳性率自相关分析有直线趋势($P<0.05$, Fig 1). 其自回归方程为 $y=1.03x$($R=0.976$, $P<0.05$),其中,y:预测下一年的值,x:当年实测值,用该方程计算 2002 年预测值为 32.42%.

万方数据

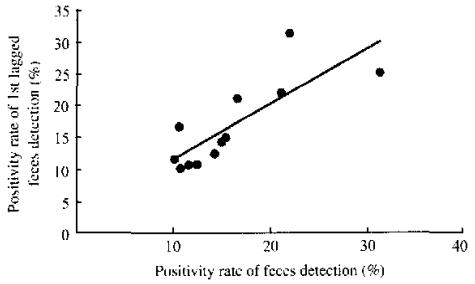

图1　粪检阳性率(%)一阶滞后散点图

Fig 1　1st lagged scatter plot of prevalence

2.1.4　用 ARIMA 模型建模预测　进行分析时,首先要确定 ARIMA 模型各阶数,主要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 Tab 2, Fig 2, 3 ARIMA(1,2,2)较好,确定值用拟合优度表示,其统计量包括标准误(standard error)、对数自然函数值(log likelihood)、AIC、SBC[5]. 将各阶数代入基本公式可得到预测方程. 预测方程为 $\hat{y}_t = 1.2877Y_{t-1} + 0.4246Y_{t-2} - 0.7123\ Y_{t-3} - 0.4305e_{t-1} + 0.9981e_{t-2} - 0.0544$. $\hat{y}_t$:预测当年的值,Y_{t-1}、Y_{t-2}、Y_{t-3}分别为上一年的实测值,余类推. $e_t - 1$为上 1 年预测值的误差,余类推. 可得该方程 2002 年粪阳率预测值为 30.73%.

表2　ARIMA 法不同模型拟合优度的比较

Tab 2　Comparison of goodness fit in different ARIMA model

	Number of residuals	Standard error	Log likelihood	AIC	SBC
ARIMA(2,1,2)	12	3.7744	-31.4755	72.9509	75.3754
ARIMA(1,1,2)	12	3.1635	-30.1229	68.2459	70.1855
ARIMA(2,2,1)	11	4.0613	-29.8506	67.7012	69.2928
ARIMA(2,2,2)	11	3.5418	-28.8566	67.7131	69.7026
ARIMA(1,2,2)	11	3.3566	-29.2258	66.4516	68.0432

ARIMA: Autoregressive integrated moving average model.

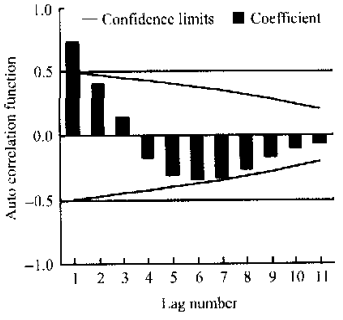

图2　自相关分析图

Fig 2　Auto correlation function

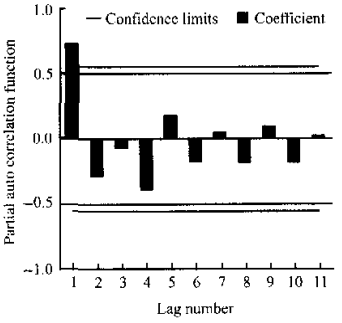

图3　偏自相关分析图

Fig 3　Partial auto correlation function

2.2　4 种方法预测效果比较　通过计算 1993-2001 年预测年平均相对误差(%)和误差平方和来判断预测效果优劣,结果如 Tab 3 示. 相对误差计算公式:相对误差 = 预测值与实测值差值的绝对值/实测值 × 100%.

表3　1993/2002 年不同方法预测效果比较

Tab 3　Predictive effect of different methods from 1993 to 2002

	Average relative error(%)	Sum of square error
Exponential smoothing	17.60	190.42
Moving average	19.74	233.86
Autoregressive model	17.76	187.29
ARIMA	15.51	100.16

ARIMA: Same as in Tab 2.

3　讨论

公共卫生事件中及时、快速地预警能够避免我国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[6]等用 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 1℃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[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. 国内亦有人运用时间序列分析对恶性肿瘤、麻疹、猩红热、乙型脑炎、尘肺、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

万方数据

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.

时间序列分析中4种方法各有特点. 移动平均法有两个优点:① 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;② 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:① 必须有N个过去观察值,如本例必须有连续3年资料;② 过去观察值中权数设置都相同,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,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,本研究结果也验证了这一点. 但有时在不同的应用条件下,模型的选择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、患者人数将逐渐增加、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.

我们曾对血吸虫病作过研究[16,17],也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[18],发现了血吸虫病及活螺密度退出还湖前后的变化趋势,但不能进行定量预测,本研究则为定量预测提供了有效工具. 时间序列模型各方法预测值的偏差大小受数据本身特点、样本量大小等因素影响. 本研究中1993/2002年的年平均相对误差和误差平方和以ARIMA模型为小,结论为在集成试点1990/2002年发病预测中,ARIMA模型预测效果较好.

【参考文献】

[1] Hong QB, Zhou XN, Sun YP. Impact of global warming on the transmission of schistosomiasis in China (Ⅱ): The activation and lethal hyperthermy temperature of oncomelania hupensis in laboratory [J]. *Zhongguo Xuexichongbing Fangzhi Zazhi* (*Chin J Schistosomiasis Control*), 2003,15(1):24-26.

[2] Li T, Yu BG, Dai YH. Impact and countermeasures on acute schistosomiasis transmission by Yangtze River flood [J]. *Zhongguo Xuexichongbing Fangzhi Zazhi* (*Chin J Schistosomiasis Control*), 2000;12(5):268-272.

[3] Sun YP, Zhou XN, Hong QB. Re-transmission of schistosomiasis Japonnica in marshl and of the Yangtze River I. Fluctuation and distribution of oncomelania hupensis [J]. *Zhongguo Xuexichongbing Fangzhi Zazhi* (*Chin J Schistosomiasis Control*), 2001;13(4):213-215.

[4] Zhang WT. *Statistical analysis curriculum of SPSS 11* (*Advanced*) [M], Beijing: the Beijing Hope Electronic Press, 2002:284.

[5] Xu GX. *Statistical forecasting and policy decision* [M], Shanghai: Shanghai university of Finance and Economics Press, 1998:158-162.

[6] Diaz J, Garcia R, Velazquez. Effects of extremely hot days on people older than 65 years in Seville (Spain) from 1986 to 1997 [J]. *Int J Biometeorol*, 2002;46(3):145-149.

[7] McCleary R, Chew KS, Merrill V. Does legalized gambling elevate the risk of suicide? An analysis of U.S. counties and metropolitan areas [J]. *Suicide life Threat Behav*, 2002,32(2):209-210.

[8] Clancy L, Goodman P, Sinclair H. Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study [J]. *Lancet*, 2002;360(9341):1210-1214.

[9] Wu JJ, Guo H, Su RH. Analysis and forecast of incidence and mortality of nasopharynx cancer by time series in Zhongshan city [J]. *Zhonghua Yiyuan Tongjixue Zazhi* (*Chin J Stat Hosp*), 2001;8(1):16-19.

[10] Wei KR, Liu Q, Wang DK. Incidence trend and prediction of nasopharyngeal carcinoma in Zhongshan during 1970-1999 [J]. *Cancer*, 2001;20(10):1065-1068.

[11] Wang WY, Chen Z. Study on the prediction of pneumoconiosis by time sequence analysis [J]. *Zhonghua Zhiye Yixue Zazhi* (*Chin J Prof Med*), 2001;28(2):27-29.

[12] Chen YK, Zeng G. Zeng-Ding phenomenon: Further demonstration and studies on its predictive value in epidemic of measles and scarlet fever [J]. *Zhonghua Liuxinghingxue Zazhi* (*Chin J Epidemiol*), 1999;20(4):200-203.

[13] Li TJ, Chen XS, Li YF. Application of the time-series method to analyse the seasonal distribution of epidemic encephalitis B incidence in Guangdong Province in the years of 1984-1993 [J]. *Zhonghua Liuxingbingxue Zazhi* (*Chin J Epidemiol*), 1998;19(2):103-106.

[14] Wang Q, Xiao YF, Jiang CD. The research of predictive methods in the prevention process of schistosomiasis epidemic trend [J]. *Shiyong Jishengchongbing Zazhi* (*J Parasitic Dis*), 1999;7(3):120-121.

[15] Xiao YF, Wang Q, Wei JB. The prediction of schistosomiasis epidemic trend in Sichuan Province from 1999 to 2001 [J]. *Shiyong Jishengchongbing Zazhi* (*J Parasitic Dis*), 2000;8(2):53-55.

[16] Zhang B, Zhang ZY, Xu DZ. Relationship between schistosomiasis prevalence and snail status in Jiangning county in Jiangsu province from 1990 to 1999 [J]. *Di-si Junyi Daxue Xuebao* (*J Fourth Mil Med Univ*), 2002;23(11):1023-1025.

[17] Zhang ZY, Xu DZ, Zhou XN. Application of LANDSAT ETM+ images in the surveillance of marshland habitat of oncomelenia snails in Jiangning county [J]. *Di-si Junyi Daxue Xuebao* (*J Fourth Mil Med Univ*), 2003;24(2):139-142.

[18] Sai XY, Cai KP, Xu DZ, Yan YP, Zhang ZY, Li YS, Zhou XN. Analysis of relationship between schistosomiasis prevalence and snail status during the period from 1990 to 2002 in the areas of "breaking dikes or opening sluice for waterstore" in Dongting Lake [J]. *Di-si Junyi Daxue Xuebao* (*J Fourth Mil Med Univ*), 2003;24(20):1878-1880.

编辑 王雪萍

万方数据