·研究原著·

文章编号:1000-2790(2006)17-1603-03

五种预测方法在退田还湖区血吸虫病发病的拟合效果评价

赛晓勇[1],邢秦菊[2],孟定茹[3],贾玉然[4],蔡凯平[5],李岳生[5],周晓农[6] ([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]解放军第323医院信息科,[3]总后勤部西安第一干休所,陕西 西安 710054,[4]兰州军区西安小寨干休所,陕西 西安 710061,[5]湖南省血防所,湖南 岳阳 414000,[6]中国疾病预防控制中心寄生虫病预防控制所,上海 200025)

Comparison of predicting effect of schistosomiasis prevalence by 5 statistical models in the areas of "breaking dikes or opening sluice for water store" in Dongting Lake

SAI Xiao-Yong[1], *XING Qin-Jun*[2], *MENG Ding-Ru*[3], *JIA Yu-Ran*[4], *CAI Kai-Ping*[5], *LI Yue-Sheng*[5], *ZHOU Xiao-Nong*[6]

[1]Department of Epidemiology, School of Preventive Medicine, Fourth Military Medical University, Xi'an 710033, China, [2]Department of Statistics, PLA 323 Hospital, Xi'an 710054, China, [3]Xi'an First Cadre Sanatorium, General Logistics Department, Chinese PLA, Xi'an 710054, China, [4]Xiaozhai Cadre Sanatorium in Xi'an, Lanzhou Military Area Command, Xi'an 710061, China, [5]Hunan Institute of Anti-epidemic of Schistosomiasis, Yueyang 414000, China, [6]Institute of Parasitic Diseases, Chinese Center for Disease Prevention and Control, Shanghai 200025, China

【Abstract】**AIM**: To compare the predicting effect of schistosomiasis prevalence by 5 different statistical models including Moving Average, Exponential Smoothing, Autoregressive Model, Autoregressive integrated moving average model (ARIMA Model) and Grey Model in the areas of "breaking dikes or opening sluice for water store" in Dongting Lake and to provide a fitted model for local schistosomiasis preventive department. **METHODS**: The 5 different statistical models were applied to predict the schistosomiasis prevalence in some experimental sites and Error Sum of Square (ESS), Average Relative Errors (ARE), Average Errors (AR) of 5 models were compared. **RESULTS**: ESS, ARE and AR of Grey Model in Jicheng were smallest; ESS and AR of ARIMA Model in Haohou were smallest; ARE of Autoregressive Model was smallest. **CONCLUSION**: Different models fit different places. The predicting effects of Grey Model and ARIMA Model are best among the 5 models.

【Keywords】Statistical prediction; ARIMA Model; Schistosomiasis; Breaking dikes or opening sluice for water store

【摘 要】**目的**:比较移动平均法、指数平滑法、自回归法、ARIMA法和灰色预测法在退田还湖地区试点血吸虫病发病拟合效果的优劣,为当地血防部门提供较为适合的拟合方法. **方法**:应用五种方法对集成垸试点和濠口试点血吸虫病患病率建模预测并比较拟合值的绝对误差、相对误差和误差平方和. **结果**:集成垸试点灰色预测法拟合值的平均绝对误差、平均相对误差和误差平方和最小;濠口试点平均绝对误差、误差平方和以ARIMA法最小,平均相对误差以自回归法最小. **结论**:不同的拟合模型适用于不同的试点,两试点以灰色预测和ARIMA模型拟合效果较好.

【关键词】统计预测;ARIMA模型;血吸虫病;退田还湖

【中图号】R181.8 【文献标识码】A

0 引言

1998年我国开始退田还湖,使血吸虫病中间宿主钉螺孳生环境发生了变化. 在应用不同方法对血吸虫病病情预测研究的基础上,对移动平均法、指数平滑法、自回归法、ARIMA法和灰色预测法进行了比较与评价,为退田还湖区血防部门找到相对精确的定量拟合方法.

1 材料和方法

1.1 材料 收集退田还湖地区华容县的集成垸试点(双退点,即退人又退田,该垸1998年完全废弃用于泄洪)和濠口试点(单退点,退人不退田即洪水期人转移、洪水过后返回种田)1990~2003年连续粪检阳性率的病情资料. 集成垸试点退田还湖后滞留人口2600人,面积为2200万平方米;濠口试点常住人口1176人,面积为297万平方米,均为湖南省血吸虫病重灾区监测试点. 全部病情资料由湖南省血防所及华容县洪山头镇血防站和澧县小渡口血防站提供.

1.2 方法

1.2.1 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$F_{t+1}=\frac{1}{N}\sum_{t-N+1}^{t}$,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:$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots -$

收稿日期:2006-03-09; 接受日期:2006-06-22

基金项目:国家"十五"科技攻关课题(2001BA705B08)

作者简介:赛晓勇. 博士生. Tel:(029)84774871 Ext. 15 Email:saixiaoyong@163.com

$\theta_q e_{t-q}$,用自相关系数识别,它的自相关系数为 :$r_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2} & 1 \leqslant k \leqslant q \\ 0 & k > q \end{cases}$. 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 $\sum_{t=k+1}^{n}(Y_t - Y)(Y_{t-k} - Y)/\sum_{t=1}^{n}(Y_t - Y)^2$. 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. $Y = \frac{1}{n}\sum_{t=1}^{n} Y_t$,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.

1.2.2 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 $\hat{z}_{t+1} = \alpha z_t + (1-\alpha)\hat{z}_t$,α 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.

1.2.3 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 :$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \cdots \phi_p Y_{t-p} + e_t$,它的偏自相关系数满足 :$\phi_{ki} = \begin{cases} \phi_i & 1 \leqslant i \leqslant p \\ 0 & p+1 \leqslant i \leqslant k \end{cases}$. 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. $\Phi_{kk} = (r_k - \sum_{i=1}^{k-1}\Phi_{k-1,i} \times r_{k-i})/(1 - \sum_{i=1}^{k-1}\Phi_{k-1,i} \times r_i)$. k = 2,3,…式中 :$\Phi_{k,i} = \Phi_{k-1,i} - \Phi_{kk} \times \Phi_{k-1,k-I}$,i = 1,2, …, k - 1.

1.2.4 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[5].

1.2.5 灰色模型 假定给定时间数据序列 $X^{(0)}$ 有 n 个值,$X^{(0)} = \{X^{(0)}(k), k = 1,2,\cdots,n\}$,作相应的 1 阶累加序列 $X^{(1)} = \{X^{(1)}(k), k = 1,2,\cdots,n\}$,则序列 $\{X^{(1)}(k), k = 1,2,\cdots,n\}$ 的 GM(1,1)模型的白化微分方程为 :$dX^{(1)}/dt + aX^{(1)} = \mu$,式中 :a 为发展灰数 ;μ 内生控制灰数. 模型检验包括残差检验、关联度检验和后验差检验. 残差检验是按预测模型计算 $\hat{X}^{(1)}(i)$,并将 $\hat{X}^{(1)}(i)$累减生成 $\hat{X}^{(0)}(i)$,然后计算原始序列 $X^{(0)}(i)$与 $\hat{X}^{(0)}(i)$的绝对误差序列及相对误差序列. 绝对误差越小越好,相对误差一般认为小于 0.5% 为好. 关联度检验是根据 $\hat{X}^{(0)}(i)$与原始序列 $X^{(0)}(i)$的关联系数计算出关联度,当 ρ = 0.5 时一般认为大于 0.6 满意了. 后验差检验需计算原始序列的标准差 S_1 和绝对误差序列的标准差 S_2,然后计算方差比和小误差概率. 若残差检验、关联度检验和后验差检验都能通过,可以用该模型预测,否则进行残差修正.

1.2.6 方法评价 以绝对误差、相对误差和误差平方和作为评价指标.

2 结果

集成垸试点和濠口试点应用 5 种方法预测的结果分别见表 1,2 .预测效果比较见表 3.

表 1 集成试点 1993 ~2002 年血吸虫粪检阳性率观察值及拟合值比较

年度	观察值	移动平均法	指数平滑法	自回归法	ARIMA 法	灰色预测法
1993	12.56	14.8167	14.4585	14.8114	13.8135	9.0656
1994	10.83	13.5917	12.7499	12.9368	11.8465	10.3309
1995	10.23	11.9983	11.0219	11.1549	8.5658	11.7727
1996	11.70	10.8183	10.3092	10.5369	7.5348	13.4158
1997	10.74	11.0650	11.5609	12.0510	11.2887	15.2883
1998	16.78	10.9750	10.8221	11.0622	14.5768	17.4220
1999	21.28	13.9200	16.1842	17.2834	16.5138	19.8535
2000	22.11	18.0233	20.7704	21.9184	26.6901	22.6244
2001	31.48	20.9450	21.9760	22.7733	31.0048	25.7821
2002	25.40	26.6567	30.5296	32.4244	30.7261	29.3804

表2 濠口试点1994 ~2002年血吸虫粪检阳性率观察值及拟合值比较

年度	观察值	移动平均法	指数平滑法	自回归法	ARIMA 法	灰色预测法
1994	7.72	8.8167	8.4914	8.0243	8.1880	7.3879
1995	7.02	8.1833	7.7971	7.3572	7.5699	7.4127
1996	6.65	7.4867	7.0977	6.6901	6.8412	7.4377
1997	3.83	6.9517	6.6948	6.3375	6.6556	7.4627
1998	4.89	5.3017	4.1165	3.6499	2.4410	7.4878
1999	7.72	4.8300	4.8127	4.6602	6.4882	7.5130
2000	9.52	6.1283	7.4293	7.3572	8.3121	7.5383
2001	9.76	8.1483	9.3109	9.0726	9.7928	7.5636
2002	7.28	9.3400	9.7151	9.3013	9.5202	7.5891

表3 不同试点各方法拟合效果比较

方法	平均绝对误差		平均相对误差		误差平方和	
	集成试点	濠口试点	集成试点	濠口试点	集成试点	濠口试点
移动平均法	3.7037	1.8426	0.1973	0.2791	233.8632	39.8669
指数平滑法	3.3849	1.5019	0.1760	0.2401	190.4221	29.1596
自回归法	3.3394	1.3733	0.1776	0.2193	187.2860	26.6314
ARIMA 法	2.5999	1.2440	0.1551	0.2226	100.1649	22.5354
灰色预测法	2.4062	1.3819	0.1512	0.2444	89.4922	29.7198

3 讨论

在对时间序列和灰色拟合模型进行选择时,应当考虑三个主要的问题:适用性、精确性和费用. 任何一种预测方法都是建立在一定的假定条件之上的,而任何一种假定条件都难以包括现实世界中所有复杂的关系因而必须考虑适用条件[1]. 移动平均法、指数平滑法、自回归法、ARIMA 法均适用于短期拟合而灰色预测还适于中期预测. 移动平均法适用于不带季节变动的反复预测,缺点是初次选择权数费时间;指数平滑法对于有、没有季节变动的反复预测均适用,建模时间与其他方法相当;自回归法适用于残差间相互不独立,过程较 ARIMA 模型简单;ARIMA 模型适用于任何序列的发展型态但计算过程复杂、繁琐;灰色预测法适用于时序的发展呈指数趋势. 各方法精确性要通过计算误差比较评价.

目前国内外统计模型在医学领域的应用已进行了大量的研究,传染病方面也有应用. 丁守銮等[2]用所建模型对肾综合征性出血热各月发病率进行了预测,结果表明 ARIMA 是一种短期内预测精度较高的预测模型,与本研究结论一致. 张彦琦等[3]曾对对数模型、指数平滑模型和 ARIMA 乘积模型的预测结果进行分析发现对数模型、指数平滑模型和 ARIMA 乘积模型的预测平均相对误差分别为14.34%,8.14%和4.89%,从而得出 ARIMA 模型效果较好的结论. 张蔚等[4]对所研究的季节性时间序列建立了乘积 ARIMA$(0,1,1)\times(0,1,1)_{12}$模型并用预测平均相对误差进行评价发现 ARIMA 乘积模型的预测效果优于指数平滑法.

灰色预测应用研究同样较多,冯丹等[5]利用 GM(1,1)模型预测大庆市流行性脑脊髓膜炎发病率和病死率并对模型精度进行了检验,显示拟合精度高. 黄春萍等[6]应用灰色模型预测克拉玛依市肺结核发病率并与线性回归模型、指数模型、多项式模型拟合效果进行比较发现,GM(1,1)模型可以对该地区肺结核发病率进行较好的短期预测. 蔡碧等[7]为了探讨灰色系统理论对血吸虫病八项疫情指标预测的可信性,用灰色理论对血吸虫病八项疫情指标建立预测模型,并用"残差建模"提高原点精度、用"等维递补灰数动态预测"来动态地预测未来结果、引进"环境干涉因子"修正预测结果,对血吸虫病各项疫情指标进行了中长期预测. 结果显示,近期预测结果得到证实,未来预测将进一步验证.

本研究结果显示集成试点灰色预测法预测的平均绝对误差、平均相对误差和误差平方和最小;濠口试点平均绝对误差、误差平方和以 ARIMA 法最小,平均相对误差以自回归法最小. 因而可以认为不同的试点适用于不同的预测模型. 集成试点以灰色预测法效果最好,濠口点 ARIMA 法效果最好,两者比较退田还湖前后的发病预测值变化趋势结论一致.

本研究存在的问题为样本量小、应用的拟合方法本身有其固有的局限性等,拟合结果有待进一步验证.

【参考文献】

[1] 徐国祥. 统计预测与决策[M]. 上海:上海财经大学出版社,1998:158 -162.
[2] 丁守銮,康家琦,王洁贞. ARIMA 模型在发病率预测中的应用[J]. 中国医院统计,2003,10(1):23 -26.
[3] 张彦琦,黄 彦,田考聪. SPSS 在医院统计预测中的应用[J]. 中国医院统计,2002,9(3):131 -134.
[4] 张 蔚,张彦琦,杨 旭. 时间序列资料 ARIMA 季节乘积模型及其应用[J]. 第三军医大学学报,2002,24(8):955 -957.
[5] 冯 丹,罗艳侠,鲍卫华,等. 流行性脑脊髓膜炎流行特征的灰色预测模型[J]. 数理医药学杂志,2003,16(2):97 -99.
[6] 黄春萍,倪宗瑭. 灰色模型在预测肺结核发病率中的应用[J]. 现代预防医学,2002,29(6):791 -793.
[7] 蔡 碧,李建屏,任先平,等. 血吸虫病灰色预测的研究[J]. 中国血吸虫病防治杂志,2000,12(2):80 -85.

编辑 许福明