利用数学模型预测季节性传染病发病率的探讨

方 华

摘要 笔者对1981～1991年北京铁路局北京地区传染病报表进行统计分析，运用数学模型对痢疾各季度发病率进行1992年预测，通过精度分析选出较为理想的预测模型，以便于季节性传染病防治工作的利用。

关键词 数学模型 传染病 发病率

了解流行性疾病的发生与时间，季节之间的潜在关系，不仅有助于提高防疫人员的工作质量，保障广大群众身体健康，也有助于认识疾病的地区性流行趋势，为防病与提高免疫能力提供可靠的依据。笔者采用时间序列模型之季节水平模型与自回归模型对北京铁路局1981～1991年的痢疾各季度的发病率作了拟合分析，并对1992年季节时段的发病率进行预测，同时比较不同的预测模型在本文预测应用中的优缺点。Box等[1]在1968年提出的简单的积分自回归模型，与传统的时间序列模型相比，其建模方法在数学上比较完善，预测精度较高，但对于有周期性变化的数据，使用一般的自回归模型，可能由于原始数据的差分变换而使部分信息丢失，导致预测精度降低。故本文在数据处理上，考虑到上述因素，结合季节水平模型与自回归模型的长处，引用了季节指数的概念，力求达到结果准确可靠。

资料来源

本文所用资料来源于北京铁路局北京地区1981～1991年各类传染病统计报表，按目的逐年整理，计算有关统计指标，从中选出痢疾的各季度发病率作为处理对象，见表1。

表1 北京铁路局1981～1991年北京地区痢疾季度发病率（/10万）

季度	年份										
	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
I	94.61 (278.26)*	122.43 (360.09)	133.42 (392.41)	215.24 (633.06)	155.70 (457.94)	112.41 (330.62)	108.37 (318..74)	81.90 (240.88)	67.10 (197.35)	58.90 (173.24)	35.00 (102.94)
II	328.65 (382.15)	403.40 (469.07)	444.03 (516.31)	494.03 (574.45)	426.29 (495.69)	299.36 (348.10)	218.83 (254.45)	148.20 (172.33)	150.90 (175.47)	118.90 (138.26)	117.80 (136.98)
III	985.96 (448.16)	1339.00 (608.64)	959.63 (436.20)	1104.92 (502.24)	879.70 (399.90)	621.19 (282.36)	485.19 (220.54)	372.90 (169.50)	603.90 (274.50)	315.80 (143.55)	309.40 (140.64)
IV	257.16 (428.63)	247.00 (411.67)	373.84 (623.07)	333.46 (555.77)	295.07 (491.78)	157.37 (262.28)	118.85 (198..08)	96.48 (160.80)	141.90 (236.50)	84.90 (141.50)	71.90 (119.83)

* 278.26=94.61/0.34 依次类推

本文作者单位：北京铁路局中心卫生防疫站 100038

建模步骤和计算方法

1.根据表1制线图初步判断季节变化是否明显，由图1可知，每年第三季度的发病率呈上升趋势，说明有明显的季节性变化。

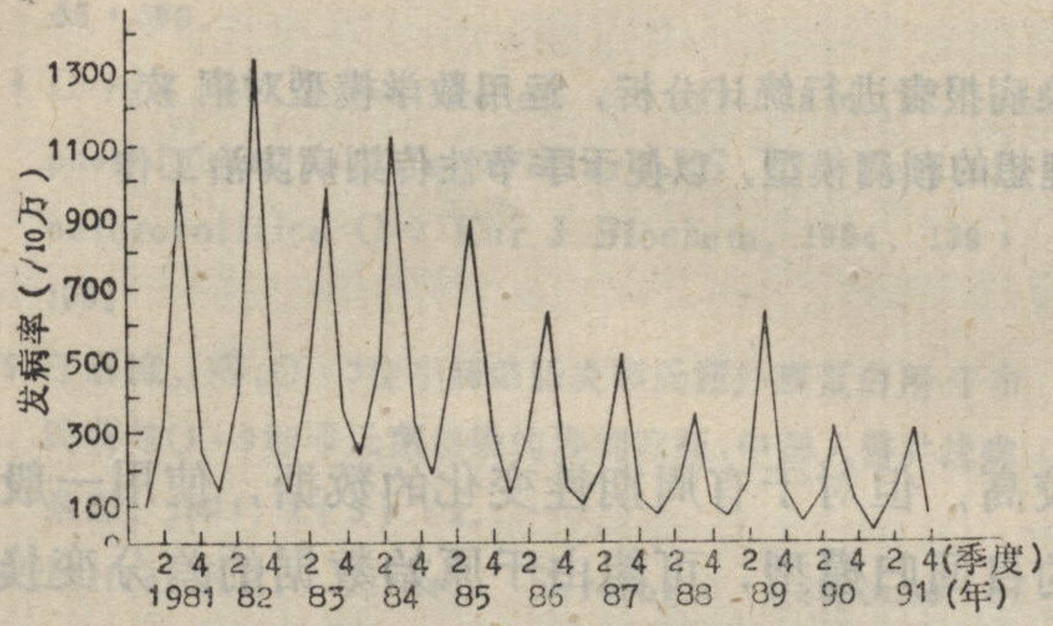

图1　1981～1991年痢疾各季度发病率

2.采用时间序列模型之季节水平模型，对痢疾各季度发病率进行预测。这种季节性水平模型适用在第t时段的期望需求量如下式：

$$\mu t=\mu Pt$$

这里的μ代表每时段平均需求量，Pt是在时段t的季节比。模型的实施分两阶段，第一阶段是用过去的需求量（X_1，X_2，… X_T）来求μ和Pt的估算值，μ在时段T的估算值是$\hat{\alpha}_T$，季节比$P_{T+\lambda}$的估算值是$\gamma_{T+\lambda}$，（$\lambda=1,2,\cdots\cdots$）。利用这些估算值求出未来时段的预测值。第二阶段为更新预测值阶段，从T时段起，每当得到一些新需求量，便可更新第一阶段的估算值。然后，利用均方拟合误差最小原则选出模型所用的平滑参数α_0和γ_0，利用α_0和γ_0所计算出来的$\hat{\alpha}_T$和$\gamma_{T+\lambda}$（$\lambda=1,2\cdots\cdots M$）来计算未来时段的预测值。其数学通式为：

$$\hat{X}_{T(\lambda)}=\hat{\alpha}_T\cdot\gamma\ (\lambda=1,2,\cdots\cdots M)\quad(1)$$

1992年各季度痢疾发病率的计算式为：

$X_{(92)}=127\cdot\gamma_{(T+\lambda)}$　　其中$\gamma_{(1-4)}$=0.34，0.86，2.2，以上过程由计算机内部完成。

3.利用Box-Jenkins普通自回归模型（AUTOREGRESSIVE MOVING AVERAGE MODEL，简记ARMA）对痢疾进行各季度的发病率预测。其数学通式为：

$$Wt=\alpha_1W_{t-1}+\alpha_2W_{t-2}+\cdots\alpha_pW_{t-p}+e_t-b_te_{t-1}-b_2e_{t-2}-b_qt-q\quad(2)$$

式中$\alpha_1\cdots\cdots$，α_p，b_1，$\cdots\cdots b_q$为参数，（p,q）为模型的阶数。

1992年各季度痢疾发病率计算式为：

$$W_{92(5)}=-1.599W_{t-1}-1.885W_{t-2}-1.731W_{t-3}-0.843W_{t-4}-0.287W_{t-5}$$

其中Wt是原始数据进行二次差分后数值，以上过程亦由计算机内部完成。

4.用趋势直线预测法和趋势季节模型预测法计算1981～1991年的季节指数[2]，结果是第一、二、三、四季度痢疾的季节指数为0.34、0.86、2.20、0.50。

然后将原始数据除以季节指数，以消除季节因素对实测资料的影响，结果见表1。若数据仍非平稳，则计算机自动对原始数据进行差分分析，最终建立季节自回归模型。其数学通式为：

$$W_t/S_t=\alpha_0+\alpha_1W_{t-1}/S_{t-1}+\alpha_pW_{t-p}/S_{t-p}\quad(3)$$

式中：α_0，$\alpha_1\cdots\cdots\alpha_p$为参数，$S_t$为季节指数。

1992年各季度痢疾发病率计算式为：

$$W_{92(2)}=-0.397W_{t-1}/S_{t-1}-0.232W_{t-2}/S_{t-2}$$

式中：W_t是原始数据进行一次差分后数值，S_t为季节指数。

从图2可知，消除季节因素的影响后，发病率呈下降趋势，即数据非平稳化，需进行差分分析。普通自回归和季节自回归模型的选择根据已给的原始数据，利用AIC准则定阶数，使AIC值最小。确定原始序列在自回归模型中选出最优拟合模型的计算程序。计算的全部过程由计算机完成。模型的识别阶数是根据时间序列的自相关函数和偏相关函数来决定模型的类型及阶数的。

结果比较与精度分析

1.表2所示为季节水平模型（Ts）、普通

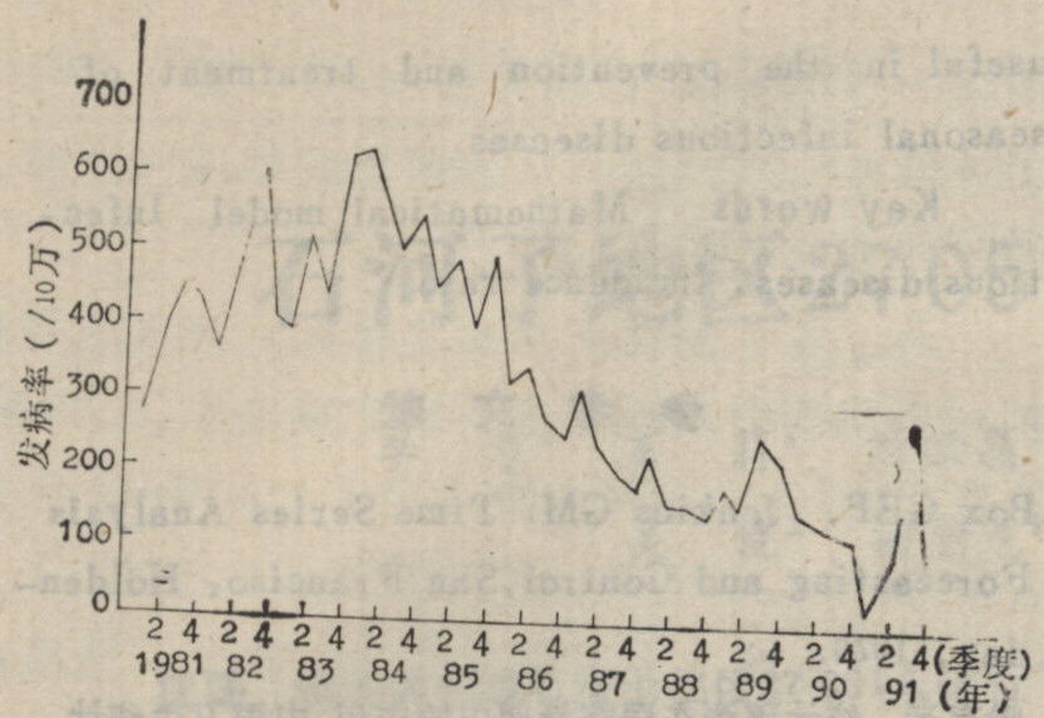

图2 消除季节因素影响后的痢疾各季度发病率

自回归模型（AR）以及季节自回归模型（SAR）三种方法对1992年各季度发病率的预测值。

2.精度分析：利用各模型计算出的1992年预测值和实测值求相对误差，以比较各种方法的预测精度。相对误差的计算方法如下[3]：

相对误差（RE）=（预测值－实际值）/实际值×100%

现将1992年各季度痢疾发病率的实际值、预测值以及相对误差列表，见表3。

表2 三种预测方法的计算结果

预测方法	1992年 痢疾 I	II	III	IV
Ts	43.49	110.00	282.66	76.74
AR	44.32	124.56	264.61	80.66
SAR	41.76	104.94	259.58	69.22

注：表中数据单位为/10万

表3 1992年1～4季度痢疾发病率预测值、实际值及相对误差

1～4季度预测值				1～4季度实际值				1～4季度相对误差				总RE
1	2	3	4	1	2	3	4	1	2	3	4	(%)
Ts 43.49	110.00	282.66	76.74					0.07	0.14	−0.01	0.06	26
AR 44.32	124.56	264.61	80.66	40.6	96.6	286.7	72.0	0.09	0.29	−0.08	0.12	42
SAR 41.76	104.94	259.58	69.22					0.03	0.09	−0.09	0.04	7

注：表中数据单位为/10万

讨 论

用时间序列模型或普通自回归模型进行预测，数据信息的准确保留是提高预测精度的关键因素。本文在比较不同方法的同时，也注意观察了对数据差分分析后的预测精度。从表3可知，采用季节性水平模型和普通自回归模型所做预测的相对误差为26%和42%，而季节自回归模型预测效果较为满意，误差仅为7%，较前两种小。其原因为前两种方法没有考虑时间序列中的季节变动的特点，致使预测值呈上升或下降的趋势，这就有可能出现与实际情况不相符的结果。而后一种方法运用季节指数以消除季节因素的影响，克服上述之不足，从而使预测精度大大提高。当然，季节自回归模型并非唯一预测季节性传染病的方法，但从其预测的结果看，还是值得推荐的。

季节性传染病发病率的预测是一个复杂的问题，其变动受到很多内、外因素的影响，但如果选用较好的数学模型，并在一定的误差范围内得到较为可信的结果，还是科学可信的。可是，当前建模型的过程虽然由计算机完成，但有很多步骤仍需人为选择，甚至需要手工计算，在基层卫生防疫部门推广应用受到一定程度的限制。但随着微机在预防预测方面的广泛应用，这些问题将会迎刃而解。

科学、准确、可靠的预测结果不仅为防疫工作提供信息，也为领导者做决策提供依据。在今后如何运用预测方法上，还需要在实践中逐步摸索、探讨，力求建模预测工作在防病治病领域真正达到“大众化”。

（北京铁路局中心卫生防疫站 王新云 副主任

医师、潘小秋医师给予了大量的支持和协作，深表谢意）

Application of the Mathematical Model to Forecast the Incidence Rates of Seasonal Infectious Diseases, *Fang Hua., Beijing Railway Hygiene and Epidemic Prevention Station, Beijing* 100038

The incidence rates of infectious diseases were selected and analysed according to data from the Beijing Railway Area during 1981～1991. We put forward the mathematical model to forecast the incidence rate of dysentery each quarter in 1992. The best mathematical model was selected from analysis of precision, and very useful in the prevention and treatment of seasonal infectious diseases.

Key words Mathematical model Infectious diseases Incidence rate

参 考 文 献

1 Box GEP, Jenkins GM. Time Series Analysis Forccasting and Control, San Franciso: Holden-day, 1970.

2 腾存远. 统计预测在医院管理中的应用. 中国卫生统计, 1988, 5(1): 20.

3 杨瑞璋, 等. 用季节自回归模型SAP(p)预测药品销售量. 中国卫生统计, 1990, 7(3): 9.

（收稿：1993-01-30）

六种不同方法预测钩体病流行强度的比较研究

张代贵[1]　王正仪[2]　羊衍惠[2]　彭国珍[2]　李世贵[3]

我们用符合正态分布的广安县1973～1989年钩体病对数发病率（LogY）与影响因素（Xi）用GM（1，1）灰色数列模型、最小二乘法、自身回归法、逐步回归、主成分、Q型聚类分析法进行了预测钩体病流行强度的方法学比较研究。样本n＝17，自变量（Xi）＝22个，并用SPSS软件包计算。计算时，逐步回归取$F^{*}_{0.2(8.8)}=1.856$，主成分取累计贡献率为90.6%，聚类参数Beta取0.67并在聚成七类后选因变量与自变量的相关系数r的绝对值最大者作为优化影响因素。继将三种多因素分析结果用钩体病流行病学观点进行定性分析以选出较优预测方法，最后再将选出的较优预测法进行定量分析、比较、验证。

结果：三种单因素预测均无显著性意义（$r\leq 0.444$，$P>0.05$）而无实用性，GM（1,1）灰色数列模型也不适宜波动幅度较大的钩体病的预测，广安县1987年就因用自身回归预测失败，发病35 985人，使1985年就曾发出的警报化为泡影，教训深刻；在定性分析中，虽然三种多因素分析均可用较少的综合指标来代替原来的较多的复杂指标，但却以逐步回归分析结果更符合钩体病流行病学观点，即抗体、鼠密度、绵雨日、雨量、疫区人口、稻田面积等均进入了模型，与传统研究认为的结果基本一致。用逐步回归模型$\hat{y}_r=-0.836+0.044x_4-0.197x_9+1.309x_{10}+0.141x_{18}-0.026x_{14}-0.026x_{16}-0.046x_{20}$，$R=0.9668$，$S=0.2977$，$F=18.38$，$P<0.001$）进行预测，$\hat{y}_r$经眉山县1988～1990年、垫江县1980～1982年资料验证，符合率在67.6%～69.8%者，占验证年份的2/3；定性分析中，主成分有6/7的自变量与气象有关，聚类有3/7的自变量与气象有关和该县以圈养猪为主，鼠是主要传染源而猪不属于主要传染源，但猪头数这一自变量却进入了聚类结果，这些均与客观不尽一致；接种菌苗人份（X_1）进入了主成分和聚类分析结果，但意义不及逐步回归结果中的抗体阳性率；n在本研究中没有大于i 5～10倍，但逐步回归分析结果仍较理想，因此赞同黄正南教授在《医用多因素分析及计算机程序》中指出的n：i为5～10：1是"很主观"的看法并推荐以逐步回归法用于钩体病的实际预测并应站在总体角度评价其运算结果而不应再用模型中单一因素来分析评判。$\hat{y}_r$也需在实践中进一步检验、完善、提高。

（本文承蒙华西医科大学公共卫生学院卫生统计教研室康春阳老师审稿并提出宝贵意见，特此致谢）

（收稿：1992-09-07　修回：1993-02-20）

1.四川省广安县计划生育委员会 638550　2.四川省卫生管理干部学院　3.广安县卫生局