灰色预测方法在疾病预测中的应用

中国人民解放军第四军医大学 流行病学教研室 汪爱勤
统计学教研室 鱼 敏

提要 本文介绍了灰色系统GM（1,1）模型，并应用此模型对恶性肿瘤死亡率资料做了预测分析。经拟合与外推预测，其结果较满意。拟合与外推预测的平均误差分别占实测值均数的1.8%和2.8%。由于此模型具有所需样本量小、无需典型的概率分布、计算简便和预测效果好等优点，可作为疾病监测的有用工具。

关键词 疾病预测 灰色动态模型

随着预防医学的发展，许多预测方法逐渐被用于疾病预测中去。但是，如何选择良好的预测模型，仍是有待解决的问题。传统的预测方法，大都建立在数理统计基础上，因而常需大量的样本和典型的概率分布，这些条件在实际中往往难以满足。为研究适合一般防疫部门的疾病预测方法，本文应用了灰色系统GM（1，1）模型对疾病预测做一尝试。此模型是一新型的预测方法，它克服了传统预测方法的上述局限性，易于被防疫人员所掌握，为疾病监测的有用工具。

模 型

一、累加生成：目的是对原始数据进行处理，使其随机性弱化和规律性强化。其定义为：

$$y(t)=\sum_{i=1}^{t} X(t) \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

式中：y(t)为累加生成数据

t为时间（1、2……N）

X(t)为数值（率）

二、均值生成：对累加生成数据y(t)按（2）式作均值生成：

$$Z(t)=\frac{1}{2}y(t)+\frac{1}{2}y(t-1) \cdots\cdots (2)$$

三、建立GM（1，1）模型，

1.建立y（t）的一阶线性微分方程：

$$\frac{dy(t)}{d(t)}=ay(t)=\mu \cdots\cdots\cdots\cdots (3)$$

我们称式（3）为GM（1，1）模型。其中a、μ为待求系数。

从方程（3）中解出a和μ满足下列方程：

$$y(t+1)=\left[X(1)-\frac{\mu}{a}\right]e^{-a(t-1)}+\frac{\mu}{a} \cdots\cdots (4)$$

2.根据最小二乘法估计参数向量，并由矩阵运算得其表达式为：

$$\hat{a}=\left\{(N-1)\left[-\sum_{t=z}^{N} X(t)Z(t)\right]+\left[\sum_{t=2}^{N} Z(t)\right]\left[\sum_{t=2}^{N} X(t)\right]\right\}\Big/ D \cdots\cdots (5)$$

$$\hat{\mu}=\left\{\left[\sum_{t=2}^{N} Z(i)\right]\left[-\sum_{t=2}^{N} X(t)Z(t)\right]+\left[\sum_{t=2}^{Z} Z^2(t)\right]\left[\sum_{t=2}^{N} X(t)\right]\right\}\Big/ D \cdots\cdots (6)$$

其中$D=(N-1)\left[\sum_{t=2}^{N} Z^2(t)\right]-\left[\sum_{t=2}^{N} Z(t)\right]^2 \cdots\cdots\cdots\cdots (7)$

3.将a和μ的估计参数$\hat{a}$和$\hat{\mu}$代入公式（4）得累加数据的预测方程。

模型的应用

一、资料来源：采用陈建国等作的《几种

主要恶性肿瘤三间分布的流行病学研究》〔1〕，其中对1973～1980年江苏省启东县恶性肿瘤死亡率进行拟合，并对1981～1984年恶性肿瘤死亡率做外推预测。

二、建立模型：原始资料见表1。

1.累加生成：按公式（1），对表1数据进行累加生成。如 $y_{(2)}=\sum_{i=1}^{2} X_{(t)}=108.46+125.87=234.33$，依此类推得表1累加生成数据。

2.均值生成：对$y_{(t)}$按公式（2）作均值生成。如：t=2时，$Z_{(2)}=\frac{1}{2}y_{(2)}+\frac{1}{2}y_{(2-1)}=171.39$，由此得$y_{(t)}$数据的均值生成数据。

3.建立GM（1,1）模型：为计算$\hat{a}$和$\hat{\mu}$，

表 1　1973～1980年启东县人口恶性肿瘤死亡率（／10万）

t	年份	死亡率 X(t)
1	1973	108.46
2	1974	125.87
3	1975	144.08
4	1976	127.45
5	1977	132.84
6	1978	135.55
7	1979	139.87
8	1980	134.03

将表1及均值生成数据列成表2。

将表2的最后一列数据代入公式(5)、(6)、(7)：

表 2　X(t)、Z(t)、$Z^2(t)$、X(t)、Z(t)数据汇总

(1) t	2	3	4	5	6	7	8	$\sum_{t=2}^{z}$
(2) X(t)	125.87	144.08	127.45	132.84	135.55	139.87	134.03	939.69
(3) Z(t)	171.39	306.37	442.14	572.28	706.48	844.18	981.13	4023.97
(4) $Z^2(t)$	29374.53	93862.58	195487.78	327504.40	499113.99	712639.87	962616.08	2820599.23
(5) X(t)Z(t)	21512.86	44141.79	56350.74	76021.68	95763.36	118075.46	131500.85	543426.74

$$D=(8-1)(2820599.23)-(4023.97)^2$$
$$=3551860.05$$

$$\hat{a}=[(8-1)(-543426.74)+4023.97\times 939.69]/3551860.05$$
$$=-0.00639$$

$$\hat{\mu}=[4023.97\times(-543426.74)+2820599.23\times 939.69]/3551860.05$$
$$=130.567$$

$$\hat{\mu}/\hat{a}\ \frac{130.567}{-0.00639}=-20433.020$$

将a和μ代入公式（4）得预测方程为：

$$\hat{y}_{(t)}=(108.46+20433.02)e^{0.0639(t-1)}-20433.02$$
$$=20541.48\ e^{0.00639(t-1)}-20433.02 \quad \cdots\cdots(8)$$

三、预测：

1.原始资料的拟合：

①利用所得方程，求出相应的累加估计值$\hat{y}_{(t)}$。$\hat{y}_{(2)}=20541.48\ e^{0.00639\times(2-1)}-20433.02=240.14$

依此类推得表3第（2）行。

②利用$\hat{y}_{(t)}$和$y_{(t)}$来估计原始数据的理论值$\hat{x}_{(t)}$。

$$\hat{X}_{(t)}=\hat{y}_{(t)}-y_{(t-1)} \quad \cdots\cdots(9)$$

如本例$\hat{x}_{(2)}=\hat{y}_{(2)}-y_{(1)}=131.68$

此外，当作外推预测时，式（9）应改为

$$\hat{X}_{(t)}=\hat{y}_{(t)}-y_{(t-1)}$$

依此类推得表3第（4）行。

从表3中可看出，理论值$\hat{x}_{(t)}$与实测值$x_{(t)}$之误差最大值为5.81，仅为实测值的4.6%，平均误差为2.46，占实测值均数的1.8%。所以，原始资料的拟合是好的。

2.模型外推预测：利用公式（8）、（9），对1981～1984年启东县恶性肿瘤死亡率做外推预测：

表 3　　理论值$\hat{X}(t)$和实测值X(t)的比较

(1)	t	2	3	4	5	6	7	8
(2)	$\hat{Y}(t)$	240.14	372.67	506.04	640.27	775.36	911.32	1048.14
(3)	Y(t)	234.33	378.41	505.86	638.70	774.25	914.12	1048.15
(4)	$\hat{X}(t)$	131.68	138.34	127.63	134.41	136.66	137.07	134.03
(5)	X(t)	125.87	144.08	127.45	132.84	135.55	139.87	134.03
(6)	$\lvert\hat{X}(t)-X(t)\rvert$	5.81	5.74	0.18	1.57	1.11	2.80	0.00

1981年即t＝9，代入（8）式：

$$\hat{y}(9)=20541.48\ e^{0.0639\times 8}-20433.02$$
$$=1185.84$$

再将$\hat{y}(t)$和$\hat{y}(t-1)$代入（9）式得$\hat{X}(t)$；

$$\text{如}\hat{X}(10)=\hat{y}(10)-\hat{y}(9)=1324.43-1185.84$$
$$=138.59$$

依此类推得1981～1984年的预测值（表4）。

表 4　　模型外推预测值与实测值比较

年份	1981	1982	1983	1984
t	9	10	11	12
$\hat{X}(t)$	137.71	138.59	139.48	140.37
X(t)	136.97	138.14	142.42	152.36
$\lvert\hat{X}(t)-X(t)\rvert$	0.74	0.45	2.94	11.99

从表4所列数据可见，除最后一个外推预测外，原始资料的拟合与近期外推预测误差都较小，外推预测的平均误差为4.03，占实测值均数的2.8%。因此认为灰色系统的GM（1,1）模型在疾病预测中是可行的，但在做远期外推预测时，需慎重。

讨　论

GM模型（Grey Dynamic Model）是通过时间序列的研究去寻找和发现事物发展变化的连续的或离散的未来时间序列，从而分析事物发展变化的规律[2]。GM（1,1）模型是灰色动态模型中最基本的，以往多用于工农业以及经济领域的预测中[3]。其括号内的第一个1，是微分方程的阶数，第二个1是变量数。本文应用灰色系统理论的GM（1，1）模型，对疾病预测问题进行了初步探讨。用传统的数理统计模型进行疾病预测时，建立方程所需样本量大，计算复杂，并受资料概率分布的限制，使其实用性和可行性受到影响[4]。而本文介绍的GM(1，1)模型所需数据少，计算简便易掌握，读者只需根据公式（1）、（2）直接列成表2形式，然后按（5）、（6）、（7）式计算得GM（1，1）模型，最后再按（8）、（9）式计算得预测值。

本文用江苏省启东县恶性肿瘤死亡率进行拟合与外推预测，推算所得理论值与实际值基本吻合，拟合良好。此外，我们用此模型对某地肺结核与细菌性痢疾的发病率[5]也做了拟合，拟合效果均较好，其中恶性肿瘤及肺结核的拟合效果最优，菌痢稍差。预测结果说明如能适当地选择该模型，预测结果会较好（图1、2、3）。

由于此模型较简单，对影响因素考虑较少，我们认为本模型主要适用于慢性疾病、肿瘤及流行因素较稳定的疾病，而对一些流行因素变化较大，或采取了新的防疫措施的疾病，其适用性有待进一步研究。此外，在做外推远期预测时需谨慎。

目前，预测分析的方法较多，每种方法各有优劣。本文介绍其中之一，使用时可根据资料的性质与作者的意图选择适当的方法或采用几种方法相互补充，使结果更可靠。

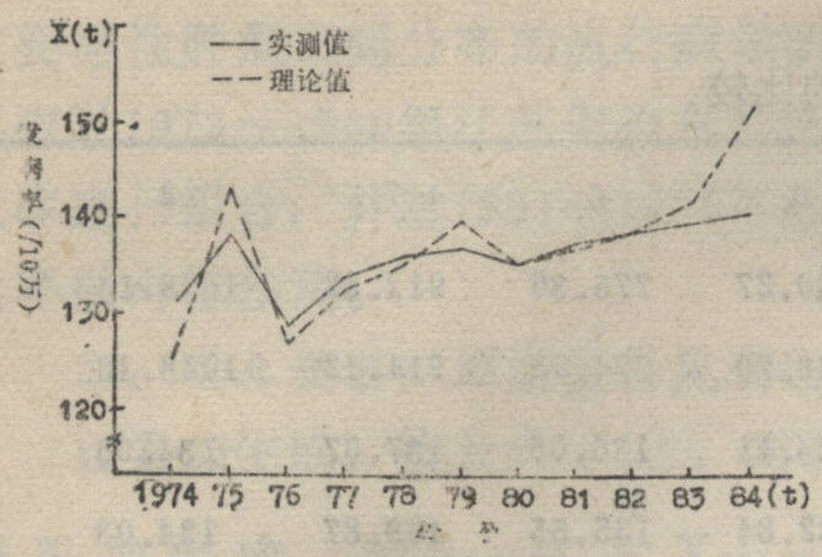

图1 1974～1984年启东县恶性肿瘤实测值和理论值 比较

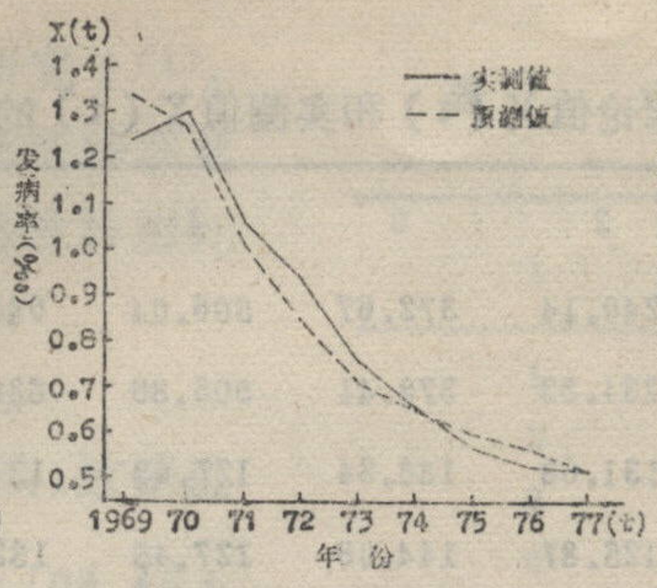

图2 1969～1977年某地肺结核实测值与预测值比较

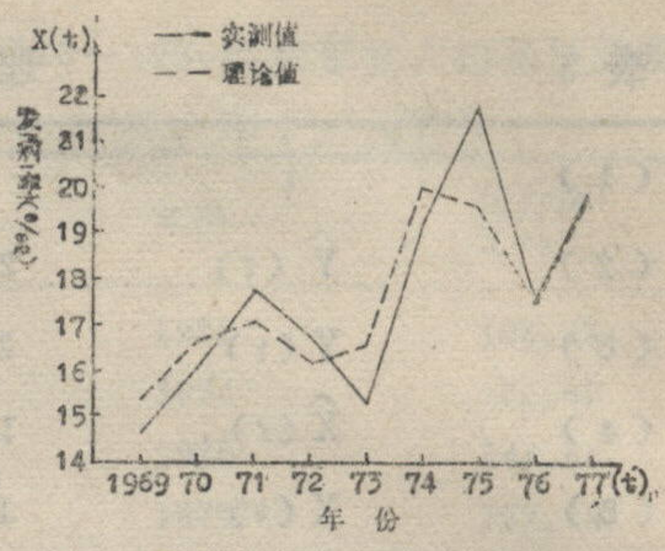

图3 1969～1977年某地细菌性痢疾实测值与理论值比较

The Application of Grey Dynamic Model in the Disease Prediction *Wang Aiqin, et al., The 4th Military Medical College, Xian*

Grey Dynamic Model was described in this paper and the death rate of malignant tumour was predicted by using this model. The data showed that the result was good for prediction. The average error of fitting and predicting were 1.8% and 2.8% of the actual mean respectively. The advantages of this model were: ①a few sample was needed; ②no need of typidal probability distribution data; ③easy counting; and④ good effect for prediction. So this model was a valuable tool, for disease surveillance.

Key words Disease prediction Grey Dynamic Model

参 考 文 献

1．陈建国，等.几种主要恶性肿瘤三间分布的流行病学研究.中华流行病学杂志 1986；7（4）：193.
2．邓聚龙.灰色系统建模思想.灰色系统.北京：国防工业出版社，1985：1.
3．赵定义，等.医院管理中的灰色预测方法.中国医院管理 1986；10（6）：14.
4．冯文权.时间序列预测技术.经济预测与经济决策技术.武昌：武汉大学出版社，1983：128.
5．王仁安，等.中国医学百科全书—医学统计学分册.上海科学技术出版社，1982：8.

（本文承郭祖超、胡琳、李良寿教授和万志恒讲师审阅，谨此致谢）

低度丝虫病流行区病人分布规律的探讨

湖北省随州市卫生防疫站 练祖银 朱家爱 张 勇 赵 华 熊天寿

我国已证实日本血吸虫中间宿主——钉螺以及血吸虫病人的概率分布规律近似负二项分布。也有报告提出丝虫病具有家庭多发特点，但用统计学方法对丝虫病家庭聚集性进行拟合，报道不多。本文就万店区有代表性的4个自然村丝虫病患者的分户分布进行普哇松和负二项分布拟合进行探讨。

一、研究对象：1985年10月对万店区33,753名社员进行丝虫病普查，查得感染率为1.02%（343/33,753），系班氏和马来丝虫混合流行区。选择了其中有代表性的4个自然村共1,175户，5,434人，先将其丝虫感染人数以户为单位进行统计，再在这基础上抽取户内人口数为5±1人的作为研究对象，计681户，3,361人，其中有丝虫病感染者74户，86人。

二、拟合结果：将丝虫感染人数按户频数分别与负二项分布、普哇松分布作配合适度检验，结果显示与负二项分布拟合较好，但与普哇松分布拟合不好（附表）。故理为低度丝虫病流行区的病人分布近似负二项分布，也就是其病人分布有家庭聚集性趋势。

附表 万店区夏家湾等村丝虫病人按户分布与负二项分布及泊松分布拟合结果

户内病人数	实际户数	负二项分布拟合之理论户数	泊松分布拟合之理论户数
0	607	606.98	600.17
1	64	63.81	75.80
2	8	8.69	4.79
3	2	1.28	0.30
χ^2		0.47	7.31
df		4-3=1	4-2=2
$P(\chi^2)$		$P>0.05$	$P<0.05$

$m=0.1263$，$S^2=0.1517$，$K=0.6280$

（本文承蒙同济医科大学流行病学教研室李光全老师审阅，谨此致谢）