中国医科大学学报 第37卷 第1期 2008年2月
Journal of China Medical University Vol.37 No.1 Feb. 2008

GM(1,1)灰色模型和ARIMA模型在HFRS发病率预测中的比较研究

吴伟[1],关鹏[1],郭军巧[2],周宝森[1]
(1. 中国医科大学 公共卫生学院流行病学教研室，沈阳 110001；2. 辽宁省疾病预防控制中心，沈阳 110005)

摘要：目的 对GM(1,1)模型和ARIMA模型在肾综合征出血热(HFRS)发病率预测中的效果进行比较。**方法** 利用1990–2001年辽宁省、丹东市和沈阳市HFRS的发病率分别建立GM(1,1)灰色预测模型和ARIMA模型，对建立的模型进行拟合。同时，对2002年3个地区的HFRS发病率进行预测，比较2个模型的拟合和预测效果。**结果** 针对辽宁省HFRS发病率建立的GM(1,1)模型和ARIMA模型的平均误差率(MER)分别为13.5143%、25.0814%；决定系数(R^2)分别为0.8961、0.6997。针对丹东市HFRS发病率建立模型的MER分别为19.7329%、20.6275%；R^2分别为0.8112、0.7628。针对沈阳市HFRS发病率建立模型的MER分别为15.1421%、18.0584%；R^2分别为0.8757、0.7889。**结论** GM(1,1)模型对于小样本以及隐含指数函数变化趋势的资料具有明显的预测优势，预测效果优于ARIMA模型，对解决时间序列类型的HFRS发病率等资料有很好的实用价值。
关键词： 肾综合征出血热；GM(1,1)模型；ARIMA模型；预测
中图分类号： R181.2 **文献标志码：** A **文章编号：** 0258–4646(2008)01–0052–04

Comparison of GM (1,1) Gray Model and ARIMA Model in Forecasting the Incidence of Hemorrhagic Fever with Renal Syndrome

WU Wei[1], GUAN Peng[1], GUO Jun-qiao[2], ZHOU Bao-sen[1]
(1. Department of Epidemiology, School of Public Health, China Medical University, Shenyang 110001, China; 2. Liaoning Provincial Center for Disease Control and Prevention, Shengyang 110005, China)

Abstract: Objective To compare the effects of GM(1,1) model and ARIMA model in forecasting the incidence of hemorrhagic fever with reual syndrome(HFRS). **Methods** GM(1,1) model and ARIMA model were established with the HFRS incidence of Liaoning province, Dandong and Shenyang from 1990 to 2001 and the two models were fit together. The forecast was made to the HFRS incidence in the three areas in 2002, and the effects of fitting and forecasting were compared. **Results** The MER of GM (1,1) model and ARIMA model for Liaoning province was 13.5143% and 25.0814%, and the R^2 of the two models 0.8961 and 0.6997 respectively. The MER of GM(1,1) model and ARIMA model for Dandong was 19.7329% and 20.6275%, and the R^2 of the two models 0.8112 and 0.7628 respectively. The MER of GM(1,1) model and ARIMA model for Shenyang was 15.1421% and 18.0584%, and the R^2 of the two models 0.8757 and 0.7889 respectively. **Conclusion** GM (1,1) model is superior in small samples on forecasting, and its effect is better than that of ARIMA model, so it is of practical value in dealing with time series data such as the incidence of HFRS .
Key words: hemorrhagic fever with renal syndrome; GM(1,1) model; auto regressive integrated moving average model; forecast

肾综合征出血热（hemorrhagic fever with renal syndrome，HFRS）主要是由某些鼠类携带传播布尼亚病毒科汉坦病毒属中不同病毒引起的一类自然疫源性疾病，其流行广、病死率高，严重危害人民的生命和健康，属于我国重点防治的传染病之一。在疾病监测的基础上，对HFRS疫情进行科学可靠的预测预报，并且有针对性地采取灭鼠和疫苗接种措施的实现，对HFRS的科学防控具有重要的指导意义[1]。

对于时间序列类型的HFRS发病率资料，比较常用的预测模型主要有时间序列模型[2,3]和GM(1,1)灰色预测模型[4]。为了比较两个模型对于HFRS资料的预测准确性，本研究对辽宁省及省内2个具有代表性的地区的HFRS发病率分别用GM(1,1)灰色预测模型和求和自回归滑动平均（auto regressive integrated moving average，ARIMA）模型进行建模和预测，然后对所建立模型的效果进行比较。

1 材料与方法

1.1 材料

发病率资料来自于辽宁省疾病预防控制中心，选取辽宁省及丹东市和沈阳市1990–2001年的HFRS年发病率(1/10万)作为拟合样本，2002年的HFRS年发病率(1/10万)作为预测样本(表1)。其间分别收集病例19 325、3 519、2 131例，获得了准确可靠的病例资料。

收稿日期：2007–06–29
基金项目：国家自然科学基金资助项目(30771860)；(70503028)
作者简介：吴伟(1981 –)，男，硕士研究生.
通讯作者：周宝森，E-mail：bszhou@mail.cmu.edu.cn

表 1　辽宁省、丹东市和沈阳市 1990–2002 年 HFRS 发病率(1/10 万)的实测值

Tab.1　HFRS incidence of Liaoning province, Dandong and Shenyang from 1990 to 2002

年份(年)	HFRS 发病率的实测值		
	辽宁省	丹东市	沈阳市
1990	0.8628	3.3154	1.1454
1991	1.1248	5.6156	1.0460
1992	0.9753	3.9718	0.7043
1993	2.1542	5.6925	1.2164
1994	2.5030	12.5938	1.7966
1995	3.1011	9.9545	2.6245
1996	3.7591	9.0294	3.1891
1997	3.5613	10.4611	2.5230
1998	4.9213	19.8464	3.6007
1999	8.5420	23.7808	4.6376
2000	7.3168	21.6605	3.9202
2001	8.5142	21.2439	5.2531
2002	8.7689	24.3876	6.4157

1.2　GM(1,1)灰色预测模型的建立

1.2.1　GM(1,1)灰色预测模型的原理：

设原始数据序列为 $x(t)=\{x(1),x(2),\cdots,x(N)\}$ $(t=1,2,\cdots,N)$经累加后生成数据，即 $y(t)=\sum_{i=1}^{t}x(t)$，$(t=1,2,\cdots,N)$，以弱化原始时间序列的随机性和波动性。然后，对累加生成数据作均值生成，$z(t)=1/2y(t)+1/2y(t-1)$，$(t=1,2,\cdots,N)$。接着，计算 D、α、u 值：

$$D=(N-1)[\sum_{t=2}^{N}z^2(t)]-[\sum_{t=2}^{N}z(t)]^2 \quad (1)$$

$$\alpha=\{(N-1)[-\sum_{t=2}^{N}x(t)z(t)]+[\sum_{t=2}^{N}z(t)][\sum_{t=2}^{N}x(t)]\}/D \quad (2)$$

$$u=\{[\sum_{t=2}^{N}z(t)][-\sum_{t=2}^{N}x(t)z(t)]+[\sum_{t=2}^{N}z^2(t)][\sum_{t=2}^{N}x(t)]\}/D \quad (3)$$

将 α、u 代入公式（4）得到预测方程，即 GM(1,1)模型

$$y(t)=[x(1)-\frac{u}{a}]e^{-a(t-1)}+\frac{u}{a} \quad (4)$$

1.2.2　原始资料拟合：

原始资料的拟合利用所得预测方程求相应的累加估计值 $y(t)$。按公式(5)估计原始资料的理论值 $x(t)$：

$$x(t)=y(t)-y(t-1),(t=1,2,\cdots,N) \quad (5)$$

1.2.3　GM(1,1)灰色预测模型的精度检验[5-8]：

人们通常采用后验差检验法来检验灰色模型的精度。用后验差检验法时，首先分别计算原始数列 x(t)和残差数列 e(t)的方差 s_1^2 和 s_2^2，然后计算后验差比值 $C=s_2/s_1$ 和小误差频率 $p=P\{|e(t)-\bar{e}|<0.6745\ s_1\}$。模型的精度由 C 和 p 共同决定。一般地，将模型的精度分为四级。当 $C\leqslant0.35$，$p\geqslant0.95$ 时，模型的精度等级为 1 级(好)；当 $0.35<C\leqslant0.50$，$0.80\leqslant p<0.95$ 时，模型的精度等级为 2 级(合格)；当 $0.50<C\leqslant0.65$，$0.70\leqslant p<0.80$ 时，模型的精度等级为 3 级(勉强)；当 $C>0.65$，$p<0.70$ 时，模型的精度等级为 4 级(不合格)。

模型最后的精度级别为 p 和 C 两者中较低的级别。如果模型的精度级别符合要求(3 级及以上级别)，则可以用来预测。GM(1,1)灰色预测模型的预测分析使用 excel 2003 完成。

1.3　ARIMA 模型的建立

1.3.1　ARIMA 模型的基本形式：

ARIMA(p,d,q)模型中 p、d、q 分别表示自回归阶数、差分阶数和移动平均数，其数学表达式为：

$$\varphi(B)(1-B)^dX_t=C+\theta(B)a_t \quad (6)$$

式中，t 表示时间；X_t 表示时间序列；B 表示后移算子，即 $BX_t=X_{t-1}$；$\varphi(B)$表示自回归算子，表示成后移算子的多项式为：$\varphi(B)=1-\varphi_1(B)-\varphi_2(B)^2-\cdots-\varphi_n(B)^n$；$\theta(B)$表示滑动平均算子，表示成后移算子的多项式为：$\theta(B)=1-\theta_1(B)-\theta_2(B)^2-\cdots-\theta_n(B)^n$；$\alpha_t$ 表示随机误差项；C 表示常数项。

1.3.2　ARIMA 模型的建立过程：

若序列 X_t 具有线性趋势，可以通过对其进行差分将线性趋势剔除掉；若 X_t 具有 d 次多项式趋势，则可以通过 d 次差分后变成平稳序列。若序列 Xt 具有指数趋势，则可以通过取对数将指数趋势转化为线性趋势，然后再进行差分以消除线性趋势。本研究中指标的时间序列都可通过差分与对数变换结合达到平稳化的目的。对平稳时间序列，采用 Box-Jenkins 模型识别方法。所有参数估计的假设检验都要求尽量有统计学意义($P<\alpha$)。ARIMA 模型的预测分析使用 SPSS13.0 软件实现。

1.4　回代拟合和点预测的效果评价

为比较 ARIMA 模型、GM(1,1)模型两种预测方法对辽宁省及省内 2 个地区 HFRS 发病率资料的拟合效果，本研究采用平均误差率(mean error rate，MER)及决定系数(R^2)两个指标对拟合效果进行评价和比较。

$$MER=\text{平均误差绝对值/实际值的均值} \quad (7)$$

$$R^2=(SS_{实}-SS_{误})/SS_{实} \quad (8)$$

公式(8)中，$SS_{实}$为实际值的方差，$SS_{误}$为误差(残差)的方差。

对于点预测，则采用残差进行预测准确性的比

较。

2　结果

2.1　GM(1,1)灰色预测模型的参数值及回代拟合结果

辽宁省 HFRS 发病率的 GM(1,1)灰色预测模型为 $y(t)=6.9617e^{0.1894(t-1)}-6.0989$；丹东市 HFRS 发病率的 GM(1,1)灰色预测模型为 $y(t)=35.5396e^{0.1492(t-1)}-32.2242$；沈阳市 HFRS 发病率的 GM(1,1)灰色预测模型为 $y(t)=7.3784e^{0.1516(t-1)}-6.2330$。

辽宁省、丹东市和沈阳市 HFRS 发病率 GM(1,1)模型的拟合结果(表 2)。

表 2　辽宁省、丹东市和沈阳市 HFRS 发病率(1/10 万) GM(1,1)模型的拟合值

Tab.2　Fitting value of GM (1,1) model for HFRS incidence of Liaoning province, Dandong and Shenyang

年份(年)	HFRS 发病率的拟合值		
	辽宁省	丹东市	沈阳市
1990	–	–	–
1991	1.4514	5.7198	1.2075
1992	1.7540	6.6404	1.4051
1993	2.1197	7.7091	1.6351
1994	2.5617	8.9498	1.9027
1995	3.0958	10.3902	2.2141
1996	3.7412	12.0624	2.5764
1997	4.5212	14.0038	2.9981
1998	5.4638	16.2576	3.4888
1999	6.6030	18.8741	4.0597
2000	7.9797	21.9117	4.7241
2001	9.6433	25.4382	5.4973

辽宁省 HFRS 发病率 GM (1,1) 模型的 p=0.9091，模型精度为 2 级，C=0.2935，模型精度为 1 级，综合起来认为该模型的精度为 2 级；丹东市 HFRS 发病率 GM(1,1)模型的 p=0.9091，模型精度为 2 级，C=0.4200，模型精度为 2 级，综合起来认为该模型的精度为 2 级；沈阳市 HFRS 发病率 GM(1,1)模型 p=1.0000，模型精度为 1 级，C=0.3229，模型精度为 1 级，综合起来认为该模型的精度为 1 级。3 个模型的精度级别均符合要求(3 级及以上级别)，可以用来预测。

2.2　ARIMA 模型的参数值及回代拟合结果

辽宁省 HFRS 发病率的时间序列经过自然对数及两次差分的转换变为平稳时间序列，ARIMA(2,2,0) 模型为：$(1+0.956B+0.73B^2)[(1-B)^2X_t]=\alpha_t$；丹东市 HFRS 发病率的时间序列经过自然对数及一次差分的转换变为平稳时间序列，ARIMA(2,1,0)模型为：$(1+0.712B^2)[(1-B)X_t]=\alpha_t+0.19$；沈阳市 HFRS 发病率的时间序列经过两次差分的转换变为平稳时间序列，ARIMA(2,2,0)模型为：$(1+0.808B+0.791B^2)[(1-B)^2X_t]=\alpha_t$。

辽宁省、丹东市和沈阳市 HFRS 发病率 ARIMA 模型的拟合结果(表 3)。

表 3　辽宁省、丹东市和沈阳市 HFRS 发病率(1/10 万) ARIMA 模型的拟合值

Tab.3　Fitting value of ARIMA model for HFRS incidence of Liaoning province, Dandong and Shenyang

年份(年)	HFRS 发病率的拟合值		
	辽宁省	丹东市	沈阳市
1990	–	–	–
1991	–	4.0080	–
1992	1.4464	6.4278	1.0194
1993	1.0371	4.3823	0.5775
1994	2.5265	9.6134	1.4196
1995	2.6180	11.4069	1.8355
1996	5.5669	8.8064	3.3876
1997	4.2798	15.9942	3.9596
1998	4.1827	15.6992	3.2487
1999	5.4701	21.8250	4.4322
2000	8.7250	20.9126	4.5172
2001	10.0484	28.5030	4.8416

2.3　GM(1,1)模型和 ARIMA 模型拟合效果的评价和比较

GM(1,1)模型和 ARIMA 模型对辽宁省 HFRS 发病率拟合的 MER 分别为 13.5143%和 25.0814%；对丹东市 HFRS 发病率拟合的 MER 分别为 19.7329%和 20.6275%；对沈阳市 HFRS 发病率拟合的 MER 分别为 15.1421%和 18.0584%。从上述数据我们可以看到，针对辽宁省、丹东市和沈阳市所建立模型的 MER，GM(1,1)模型<ARIMA 模型。

GM(1,1)模型和 ARIMA 模型对辽宁省 HFRS 发病率拟合的 R^2 分别为 0.8961 和 0.6997；对丹东市 HFRS 发病率拟合的 R^2 分别为 0.8112 和 0.7628；对沈阳市 HFRS 发病率拟合的 R^2 分别为 0.8757 和 0.7889。从上述数据我们可以看到，针对辽宁省、丹东市和沈阳市所建立模型的 R^2，GM(1,1) 模型>ARIMA 模型。

2.4　GM(1,1)模型和 ARIMA 模型点预测结果及点预测效果的评价与比较

GM (1,1) 模型和 ARIMA 模型对辽宁省 2002 年 HFRS 发病率预测结果分别为 11.6539 和

11.9315；对丹东市 2002 年 HFRS 发病率预测结果分别为 29.5323 和 33.2897;对沈阳市 2002 年 HFRS 发病率预测结果分别为 6.3969 和 6.5062。

GM(1,1)模型和 ARIMA 模型对辽宁省 HFRS 发病率点预测的残差分别为 2.8850 和 3.1626;对丹东市 HFRS 发病率点预测的残差分别为 5.1447 和 8.9021；对沈阳市 HFRS 发病率点预测的残差分别为-0.0188 和 0.0905;从上述结果我们可以看到,针对辽宁省、丹东市和沈阳市所建立模型,GM(1,1)模型的预测准确性高于 ARIMA 模型。

3　讨论

本研究选择 ARIMA 模型和 GM(1,1)模型,是因为这两种方法是比较常用的预测模型,并且属于两种不同的预测思想和预测理论。ARIMA 模型是以微积分和数理统计等传统数学理论为基础建立起来的一大类预测模型中最为成熟的时间序列预测方法之一。HFRS 发病率水平的变化受到各种因素的影响与制约,往往呈现不规则性,在数据上则表现为趋势性和随机性的特点。此外,未来的 HFRS 发病率水平与过去和现在的 HFRS 发病率水平间具有某种内在的联系,从这个意义上说,HFRS 发病率序列属于一种马尔柯夫随机过程。ARIMA 模型建立在马尔柯夫随机过程的基础上,既吸取了回归分析的优点又发挥了移动平均的长处。它根据数据序列的自相关函数、偏相关函数建立起线性的数据间相互依赖的定量模型[9],因而反映了现在的 HFRS 发病率水平和过去的 HFRS 发病率水平之间的本质联系。在预测精度方面,ARIMA 模型对噪声进行了分析处理,只剩下当时和历史无关的白噪声,使其成为线性模型的最优预测。GM (1,1)模型是是建立在灰色理论基础上的,它基于随机的原始时间序列,按时间累加后形成新的时间序列,新序列所呈现的规律即可用一阶线性微分方程的解来逼近。它在一定程度上克服了传统预测模型多建立在数理统计基础上、并需要大量样本和典型概率分布的局限性,一定程度上有助于减少时间序列的随机性和提高预测精度。

我们根据辽宁省、丹东市和沈阳市 1990~2001 年的 HFRS 发病率资料分别建立的 GM(1,1)模型和 ARIMA 模型,然后对 3 个地区 2002 年 HFRS 发病率进行预测。从拟合和预测的效果来看,前者均优于后者。出现这种结果的原因主要是：

首先，在本研究中,ARIMA 模型表现不佳很大程度上和 HFRS 发病率资料样本量过小有关。有研究指出,n 在 20 左右时,对 ARIMA 模型已经算是小样本,这种情况下的预测误差可能会较大[10]。本研究也表明 ARIMA 模型对小样本预测的预测效果比较差。

其次,辽宁省、丹东市和沈阳市 1990~2001 年的 HFRS 发病率的时间序列隐含指数函数变化趋势，这是本研究中 GM (1,1) 模型预测效果好于 ARIMA 模型的重要原因之一。有研究证明,在 GM (1,1)模型中,经一阶线性微分方程的解逼近所揭示的原始时间序列呈指数变化规律。因此，当原始时间序列隐含着指数变化规律时,GM(1,1)模型的预测效果是令人满意的[11]。

综上所述,在实际工作中,对于像 HFRS 发病率这种小样本资料，并且隐含指数函数变化趋势时采用 GM(1,1)模型进行建模预测能获得较好的预测效果。

参考文献：

[1] 郭娜娜,李琦,张艳波,等.肾综合征出血热预测方法研究现状[J]. 现代预防医学,2006,33(6):927-929.

[2] HU W,MENGERSEN K,Bi P,et al. Time-series analysis of the risk factors for haemorrhagic fever with renal syndrome: comparison of statistical models [J]. Epidemiol Infect,2007,135(2):245-252.

[3] 郭秀华,曹务春,胡良平,等.肾综合征出血热发病率季节性时间序列预测模型 [J]. 中国人兽共患病杂志,2003,19(4):121-123.

[4] 郭秀花,曹务春,张习坦.肾综合征出血热流行病学数学模型研究进展 [J]. 中国公共卫生,2003,19(4):477-478.

[5] 任正洪.孕产妇死亡率的灰色预测 [J]. 中国卫生统计,2005,22(1):20-22.

[6] 黄春萍,倪宗瓒.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] 周诗国.我国人口的灰色预测模型研究及其应用 [J]. 数理医药学杂志,2005,18(4):307-309.

[8] 刘涛.人口死亡率的灰色预测模型[J]. 数理医药学杂志,2004,17(4):290-291.

[9] 袁振洲.应用自回归积分移动平均法预测铁路货源货流发展趋势[J]. 铁道学报,1996,18(400):52-56.

[10] 张明玉.小样本经济变量相关关系检测的数学模型 [J]. 预测,1998,17(3):57-63.

[11] 吴培乐.灰色预测法在西安市农村电话预测中的应用 [J]. 西安邮电学院学报,2002,7(2):20-23.

（编辑　孙宪民）