非统计理论在灰色建模中的应用

李 勇[1] ,冯仲科[1] ,鲁绍伟[2]
(1. 北京林业大学 测绘与 3S 技术中心 ,北京 100083 ;2. 石家庄经济学院 ,石家庄 050031)

摘 要:为了解决灰色建模过程中利用残差模型来修正原灰色模型的繁琐工作 ,提出利用稳健估计理论对灰色建模的数据序列进行分析、筛选 ,找出数据列中存在的不确定信息 ,并达到直接建立灰色模型的目的。针对数据中存在不确定信息的实际情况 ,利用非统计理论对所研究的数据序列进行分析 ,对不确定信息中是否含有误差有一个确切的结论 ,进而找到一个比较方便而准确的建立灰色模型的方法。

关键词:非统计理论 ;灰色理论 ;不确定信息
中图分类号: X824 **文献标识码**: A **文章编号**: 1005-3409 (2007) 06-0074-03

Application of the Non-statistical Theory in Building Gray Model

LI Yong[1] ,FENG Zhong-ke[1] ,LU Shao-wei[2]
(1. *Survey Mapping and 3S Technology Center , Beijing Forestry University , Beijing* 100083 , *China;*
2. *Shijiazhuang University of Economics , Shijiazhuang* 050031 , *China*)

Abstract: To solve the redundant work about using deformed model to modify grey model during building grey model. Robust estimation theory is used for finding the uncertainty information from data column of the grey theory by analyzing and filtering it. Due to the data's uncertainty information ,a specific result that whether the information exists error or not is obtained by using Non-statistical Theory ,and a method is found to build grey model conveniently and correctly.

Key words: Non-statistical Theory ; grey theory ; uncertainty information

灰色系统理论广泛存在于自然界、人类社会等领域 ,其主要特点是相对于传统的随机数理统计理论 ,对样本量及其分布没有严格的要求 ,并且运算简洁方便。灰色系统理论从 1982 年由邓聚龙创立到现在[1] ,获得了飞速的发展 ,已渗透到自然科学和社会科学的许多领域 ,不仅成功用于工程控制、社会经济和生态系统等 ,而且在一些复杂多变的系统如气象、水利等方面取得显著的成绩 ,在系统分析与建模、预测与决策和灰色控制等方面为人们提供了有效的工具。

灰色模型以 GM(1 ,1)[2-5]模型为主 ,在建模的过程中 ,主要是应用最小二乘估计 ,也就是在计算时遵循残差平方和最小的原则。由于最小二乘估计具有不能抵抗粗差 ,只是具有均衡误差的作用。因此 ,利用最小二乘估计进行灰色建模时 ,当建模的精度低时就要利用残差模型对原建模模型进行修正。也就是对于少数据建模的灰色理论 ,有些时候利用最小二乘估计必然会导致建模数据的一次性利用率低的问题。

针对上述问题提出了利用稳健估计理论 ,直接进行灰色建模。在利用稳健估计进行建模的过程中 ,建模的精度完全满足建模要求。但是 ,在利用稳健估计进行建模时 ,对于利用率低的数据 ,即在文中提到的不确定信息 ,它是否是粗差 ,是否在建模时进行必要的处理。本文也正是从这一点出发 ,利用非统计理论 (以灰色系统理论、模糊集合理论、信息熵理论、贝叶斯理论、范数理论和神经网络理论等为基础) 对该数据列进行分析 ,达到精确建立灰色模型的目的。

1 数据来源

本文的数据采用刘思峰等编写的《灰色系统理论及其应用》中 115 页例 6.4.1[6] 。原题如下:

例 6.4.1 湖北省云楚县的油菜发病率数据为

$X^{(0)} = (X^{(0)}(1), X^{(0)}(2), X^{(0)}(3), X^{(0)}(4), X^{(0)}(5), X^{(0)}(6), X^{(0)}(7), X^{(0)}(8), X^{(0)}(9), X^{(0)}(10), X^{(0)}(11), X^{(0)}(12), X^{(0)}(13)) = (6, 20, 40, 25, 40, 45, 35, 21, 14, 18, 15.5, 17, 15)$

建立 GM(1 ,1) 模型 ,得时间响应式为

$\hat{x}^{(1)}(k+1) = -567.999e^{-0.06186k} + 573.999$

表 1 误差检验

序号	原始数据 $x^{(0)}(k)$	模拟值 $\hat{x}^{(0)}(k)$	残差 $\varepsilon(k) = x^{(0)}(k) - \hat{x}^{(0)}(k)$	相对误差/ % $\Delta_k = \frac{\lvert\varepsilon(k)\rvert}{x^{(0)}(k)}$
2	20	35.6704	-15.6704	78.4
3	40	33.4303	6.5697	16.4
4	25	31.3308	-6.3308	25.3
5	40	29.3682	10.6318	26.6
6	45	27.5192	17.4808	38.9
7	35	25.7901	9.2099	26.3
8	21	24.1719	-3.1719	15.1
9	14	22.6543	-8.6543	61.8
10	18	21.2307	-3.2307	17.9
11	15.5	19.8974	-4.3974	28.4
12	17	18.6478	-1.6478	9.7
13	15	17.4768	-2.4768	16.5

作累减还原得:

$\hat{X}^{(0)} = \{\hat{x}^{(1)}(k)_2^{13}\} = (35.6704, 33.4303, 31.3308, 29.3682, 27.5192, 25.7900, 24.1719, 22.6534, 21.2307,$

收稿日期:2006-11-20
基金项目:国家自然科学基金项目(90302014) ;北京市自然基金项目(4041002)
作者简介:李勇 ,副教授 ,博士生 ,主要从事 3S 技术应用研究。
通信作者:冯仲科 ,教授 ,博士生导师 ,主要从事 3S 技术应用研究。

19.8974,18.6478,17.4768)

如表 1 所示,由于模拟误差较大,因而建立残差模型进行修正。对于此模型,对 $k = 10,11,12,13$ 等 4 个数据进行修正。修正后的精度如下表:

表 2　修正后的误差检验

序号	原始数据	模拟值	残差	相对误差/ %
10	18	17.1858	0.8142	4.5
11	15.5	16.4799	- 0.9799	6.3
12	17	15.7604	1.2396	7.3
13	15	15.0372	- 0.0372	0.2

2　利用稳健估计进行数据分析

2.1　稳健估计建模

利用 SPSS 软件得到如下结果[7~11](表 3)。

表 3　全信息迭代结果

迭代次数	残差	参数	
		a	*b*
0.1	2471.250	1.000	1.000
1.1	187.839	- 0.081	1.006
2.1	179.102	- 0.064	1.294
3.1	110.524	0.062	28.840
4.1	84.094	0.101	42.761
5.1	79.914	0.099	43.915
6.1	79.605	0.078	38.572
7.1	79.605	0.078	38.572
8.1	79.605	0.078	38.572
9.1	79.605	0.078	38.572

从表 3 可看出,共迭代了 9 次,迭代终止。残差的绝对值为 79.605。建立 GM(1,1)模型,得时间响应式为

$$\hat{x}^{(1)}(k+1) = -488.5128e^{-0.078k} + 494.5128$$

表 4　全信息的稳健估计建模精度

序号	原始数据	模拟值	残差	相对误差/ %
2	20	36.6558	- 16.6558	83.3
3	40	33.9053	6.0947	15.2
4	25	31.3652	- 6.3612	25.4
5	40	29.0080	10.9920	27.5
6	45	26.8314	18.1686	40.4
7	35	24.8181	10.1819	29.1
8	21	22.9558	- 1.9558	9.3
9	14	21.2333	- 7.2333	51.7
10	18	19.6401	- 1.6401	9.1
11	15.5	18.1664	- 2.6664	17.2
12	17	16.8032	0.1968	1.1
13	15	15.5424	- 0.5424	3.6

将稳健估计建模的精度(表 4)与原例题中的建模精度(表 1)进行比较,可以看出。在利用稳健估计进行迭代后,绝大部分数据模拟精度提高。从表 4 可以看到前面的几个数据与后面的存在明显差异,拟和误差偏大,是否前面几个数据存在系统误差,是对该数据列产生的第一个疑点。另外,对于 2 号数据,它拟和的精度非常低,个别的数据也有类似问题,是否数据中含有粗差,是对该数据列产生的第 2 个疑点。本文也正是从这两个疑点出发,利用非统计理论对数据进行分析,以便得到确切的结论。

3　利用非统计理论进行分析

3.1　检验数据是否含有系统误差

设数据序列

$X_i = \{x_i(k) \mid k = 1,2,\cdots,n\}, i = 1,2,\cdots,s$ 和

$X_j = \{x_j(k) \mid k = 1,2,\cdots,n\}, j = 1,2,\cdots,p$

研究 X_i 和 X_j 之间的系统误差。

取 X_i 序列的第一个元素 $X_i(1)$,建立参考序列

$X_0 = \{x_0(k) \mid k = 1,2,\cdots,n\}$

式中:$x_0(k) = x_i$

根据灰色关联系数的定义得:

$$\gamma(x_0(k), x_j(k)) = \frac{\Delta_{\min} + \xi\Delta_{\max}}{\Delta_{ij} + \xi\Delta_{\max}}$$

式中:ξ——分辨系数,$\xi \in (0,1)$。

$$\Delta_{ij} = \Delta_{ij}(k) = |x_0(k) - x_j(k)|$$

$$\Delta_{\min} = \min_j \min_k \Delta_{ij} \quad \Delta_{\max} = \max_j \max_k \Delta_{ij}$$

灰色关联度为

$$\gamma_j = \gamma(x_0, x_j) = \frac{1}{n}\sum_{k=1}^{n}\gamma(x_0(k), x_j(k)), j = 1,2,\cdots,p$$

定义灰色关联度的绝对差为

$$d_{ij} = |\gamma(x_0, x_i) - \gamma(x_0, x_j)|$$

式中:$\gamma(x_0, x_i)$ ——X_i 对 X_0 的灰色关联度;$\gamma(x_0, x_j)$ ——X_j 对 X_0 的灰色关联度。

系统误差的显著性诊断原理为

(1)若 d_{ij} 越大,则 X_i 与 X_j 之间的系统误差越显著。

(2)若 d_{ij} 越小,则 X_i 与 X_j 之间的系统误差越不显著。

将数据(6,20,40,25,40,45,35,21,14,18,15.5,17,15)分成 2 组,前 7 个(6,20,40,25,40,45,35)1 组,后 6 个(21,14,18,15.5,17,15)1 组。并取第 1 个数据列的第 1 个数据构成参考序列,即:

$x_0(K) = x_1(1) = 6$

求差序列为 $\Delta_i(k) = |x_i(k) - x_0(k)|$

从上述差序列中求出两级最大、最小差值,即

$$\max_j \max_k \Delta_i(k) = 39 \quad \min_j \min_k \Delta_i(k) = 0$$

取分辨系数为 0.5,计算关联系数为

$\gamma_1 = (1, 0.582, 0.364, 0.506, 0.364, 0.333, 0.402)$

$\gamma_2 = (0.565, 0.709, 0.619, 0.672, 0.709, 0.684)$

$d_{12} = |0.66 - 0.507| = 0.153$

因为关联系数很小,可以说两组数据间没有系统误差。

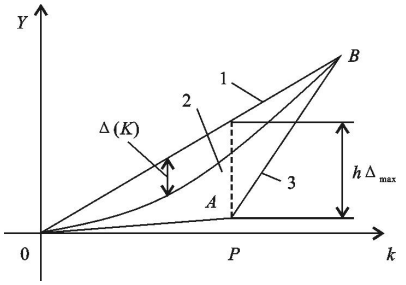

图 1　粗大误差的灰色判别原理

3.2　检验数据是否含有粗差

图 1 中横坐标 k 是测量次数,纵坐标 Y 是累加测量值,直线 1 是理想真值的一次累加生成曲线,曲线 2 是实际测量值的一次累加生成曲线,$\Delta(k)$ 是直线 1 与曲线 2 的差值。

测量值累加曲线可以用一根折线来包络,由于测量次数的中值最有可能是最大距离 Δ_{max},将测量次数的中值 p 作为包络折线的转折点。考虑到测量次数的随机变化,将最大距 Δ_{max} 离乘以 h。过测量次数为 p 的测点,将其距离值向下延伸到 $h\Delta_{max}$ 得到图中的 A 点,连接点 A、原点 O 和测量终点 B,组成一条折线,如图中的折线 3,折线 3 和参考直线 1 所围成的区域,构成了测量累加曲线的上下区间。如果有超出该范围的测量值,即认为是粗差,应予剔除。

灰色判断准则是,设有 n 个从小到大排序的测量数据

$$X^{(0)} = (x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n))$$

累加测量序列为 $(y_2(1), y_2(2), \cdots, y_2(n))$

设第 p 个测点为折线的转折点,取

$$P=\begin{cases} n/2, & n\text{为偶数} \\ (n+1)/2, & n\text{为奇数} \end{cases}$$

累加曲线的上界为参考直线 1,方程为

$$y_1(k) = \frac{y_1(n)}{n} k = \bar{x}k, k = 1, 2, \cdots, n$$

式中:$\bar{x}$——测量数据的均值。

折线 3 的方程为

$$y_3(k) = \begin{cases} \bar{x}k - h\dfrac{\Delta_{max}}{p}k, & 1 \le k \le p \\ \bar{x}k - h\dfrac{\Delta_{max}}{n-p}(n-k), & p \le k \le n \end{cases}$$

对于常数 h 习惯取 3.75,若满足

$$y_3(k) \le y_2(k) \le y_1(k)$$

则认为第 k 个数据不含有粗差。又因为 $y_2(k) \le y_1(k)$ 总是满足,故判断规则可以简化为:$y_2(k) < y_3(k)$ 成立,就认为第 k 个数据含有粗差。

通常一次只能判断一个数据,先怀疑两端的数据,若为粗差,需要将此数据剔除后,用余下的数据重新计算。就这样重复进行计算,直到把所有数据都检验完为止。

将数据(6,20,40,25,40,45,35,21,14,18,15.5,17,15)以从小到大的顺序排列为(6,14,15,15.5,17,18,20,21,25,35,40,40,45),为便于计算,把数据的参考曲线定为直线。

对上述数据有,$p = 7$, $\bar{x} = 23.962$, $\Delta_{max} = 39$。首先判断第一个数据 $y_2(1) = 6$; $y_3(1) = 23.962 - 3.75 \times 39/7 = 3.069$

因为 $y_2(1) > y_3(1)$,所以该数据不含有粗差。用类似方法,逐个检验,最后判断该数据列都不含有粗差。

3.3 对建模数据的整体分析

经过对数据系统误差和粗差的检验,我们已经对原来定义的不确定信息有了新的认识。由于数据大小而致使建模的拟和误差的大小并不能代表数据本身的精度,对于建模中的数据,既不要忽视数据的质量,同时也要注意充分地利用数据信息反映事物的本来性质。

为了能利用数据的信息,又要反映出事物的规律,采取灰色理论中的新陈代谢原理,尽量利用新的信息源,建立灰色模型。在提高建模精度的前提下,仍采用稳健估计来替代最小二乘估计。利用数据的后 6 个数据建立 GM(1,1)模型,得时间响应式为

$$\hat{x}^{(1)}(k+1) = -339.377e^{-0.061k} + 374.377$$

从表 5 中可以看出,除 2 号数据的精度相对低外,其他均满足灰色建模的要求。2 号的精度低,也反映出 2 号数据的贡献率较低。尽管有 2 号数据的影响,但对建模的整体数据而言没有太大影响,这也正体现了稳健估计的特点。

表 5 部分信息稳健估计建模精度

序号	原始数据	模拟值	残差	相对误差/%
1	21	20.0832	0.9168	4.4
2	14	18.8948	-4.8948	35.0
3	18	17.7766	0.2234	1.2
4	15.5	16.7247	-1.2247	7.9
5	17	15.7350	1.2650	7.4
6	15	14.8038	0.1962	1.3

4 结论和展望

(1)利用非统计理论对数据进行分析,对数据存在的不确定信息有了更进一步的认识。对数据的研究并没有结束,还需要更深入地加以研究。

(2)利用稳健估计理论与灰色新陈代谢原理相结合建立灰色模型,是一个较好的模式。既利用了数据信息,又考虑了数据对建模精度的影响。

(3)对于建模过程中出现的某些数据拟合精度较低的问题,是一个非常实际的问题。首先要看到问题存在的客观性,在确定了数据真实性的基础上,应该认识到该数据的重要性,这也是在解决问题中不断去发现问题的过程。只有正确对待和充分加以重视的前提下,才会尊重数据的客观事实。

参考文献:

[1] 邓聚龙.灰预测与灰决策[M].武汉:华中科技大学出版社,2002:71-96.

[2] 王中宇,夏新涛,朱坚民.非统计原理及其工程应用[M].北京:科学出版社,2005:95-98.

[3] Wang Zhongyu, Zhu Jianmin, et al. Research development of the grey error throry and the application in the dynamic measurement[C]. SPIE, 2003:447-451.

[4] 刘大杰,陶本藻.实用测量数据处理方法[M].北京:测绘出版社,2000:51-71.

[5] 傅立.灰色系统理论及其应用[M].北京:科学技术文献出版社,1992:79-80.

[6] 刘思峰,郭天榜,党耀国.灰色系统理论及其应用[M].北京:科学出版社,1999:113-116.

[7] 冯仲科,罗旭,石丽萍.森林生物量研究的若干问题及完善途径[J].世界林业研究,2005(3):25-28.

[8] 郭清文,冯仲科,张彦林,等.单木生物量模型误差分析及定权方法探讨[J].中南林业调查规划,2006(1):5-9.

[9] 罗永会,要秉文,姚少巍.Matlab 稳健回归在建立校准曲线中的应用[J].计量技术,2006(1):54-56.

[10] 张文彤.SPSS 统计分析高级教程[M].北京:高等教育出版社,2004:153-157.