稳健估计在林业灰色建模中的应用

李双柱[1], 李勇[2], 王维[3]

(1 .平朔煤炭工业公司, 山西平朔036006 ;2 .辽宁省交通高等专科学校, 辽宁沈阳110122 ;3 .沈阳煤业集团公司, 辽宁沈阳110122)

摘要　简析稳健估计理论在林业灰色建模中的应用,并将理论与实例结合对稳健估计与最小二乘估计进行比较。结果表明,利用稳健估计理论来进行林业灰色建模,不但可以抗击粗差的影响,而且预测精度基本满足在80 % 以上。

关键词　稳健估计;最小二乘估计;灰色理论

中图分类号　S11 +7　　**文献标识码**　A　　**文章编号**　0517 - 6611(2009) 01 - 00005 - 02

The Application of Robust Estimation in Forestry Gray Modeling

LI Shuang-zhu et al　(Pingshuo Coal Industry Company , Pingshuo , Shanxi 036006)

Abstract　This study aimed to analyze the application of Robust Estimation Theory in forestry gray modeling and compare the Robust Estimation and the Least Squares Estimation by combining theory with examples . The results showed that the application of Robust Estimation Theory in forestry gray modeling could not only avoid the effect of gross error , but also reach the predicted precision over 80 % .

Key words　Robust Estimation ; The Least Squares Estimation ; Gray Theory

灰色系统理论从1982 年由邓聚龙创立到现在[1],已获得了飞速的发展,并渗透到自然科学和社会科学的许多领域。灰色系统建模与统计方法有本质的区别[2 - 3], 灰色模型以GM(1 ,1) 模型[4 - 5] 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的平方和最小的原则。利用残差模型对GM(1 ,1) 模型进行修正,证明在残差序列中仍然存在有用的数据信息。笔者为解决灰色建模数据的利用率问题,研究了灰色建模的计算方法及理论。

1　灰色GM(1 ,1) 模型原理分析

1.1　GM(1 ,1) 模型原理　设$X^{(0)}$为非负序列$X^{(0)} = (x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n))$。式中,$x^{(0)}(k) \geq 0$, $k = 1,2,\cdots,n$; $X^{(1)}$为$X^{(0)}$的1 - AGO 序列$X^{(1)} = (x^{(1)}(1), x^{(1)}(2), \cdots, x^{(1)}(n))$;式中,$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i)$, $k = 1,2,\cdots,n$; $Z^{(1)}$为$X^{(1)}$的紧邻均值生成序列$Z^{(1)} = (z^{(1)}(2), z^{(1)}(3), \cdots, z^{(1)}(n))$。式中,$z^{(1)}(k) = 0.5x^{(1)}(k) + 0.5x^{(1)}(k-1)$, $k = 2,3,\cdots,n$。称$dx^{(1)}/dt + ax^{(1)} = b$为灰色微分方程$x^{(0)}(k) + az^{(1)}(k) = b$的白化方程。将灰色微分方程$x^{(0)}(k) + az^{(1)}(k) = b$写成矩阵形式是:

$$Y = Ba。$$

式中,
$$Y = \begin{matrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \cdots \\ x^{(0)}(n) \end{matrix} \quad B = \begin{matrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \cdots & \cdots \\ -z^{(1)}(n) & 1 \end{matrix} \quad a = \begin{matrix} a \\ b \end{matrix}$$

在满足$\varepsilon^T\varepsilon = \min(\varepsilon = Y - Ba)$的前提下,即利用最小二乘估计得到参数a , $a = (B^TB)^{-1}B^TY$。

1.2　对灰色GM(1 ,1) 模型的分析　从灰色GM(1 ,1) 模型利用最小二乘估计答解灰色方程的过程来看,实际是把数据序列看作是等精度序列。而对实际中的数据序列,完全符合等精度的极少,绝大多数是不等精度的。同时,对于灰色建模的数据序列而言,有的数据要跨越几年或几十年,数据序列的精度问题是不可以回避、必须解决的问题。

傅立在《灰色系统理论及其应用》中已对GM(1 ,1) 加权模型作过论述。主要是针对原始数据列中的数据精度不等,根据数据的可信度来进行赋权,可信度高的权就大,也就是让可信度大的数据在建模时的贡献率高一点。傅立介绍的加权方法是加权的最小二乘估计。在定权的时候,他提出了根据数据精度随时间成正比例变化的数据如何确定权重,符合这样假设的数据太少,但毕竟是从数据的精度出发的,有可以借鉴的地方。由于灰色建模数据序列跨越的时间各有不同,在数据的获取上人员、仪器、外界条件都不尽相同,为了解决灰色建模数据如何来确定权重这个棘手的问题,已有学者提出利用具有抗击粗差能力的稳健估计来处理灰色建模中遇到的数据序列精度不等的问题[6 - 7]。

2　稳健估计的原理

稳健估计是在最小二乘估计的基础上发展起来的。由于最小二乘估计不具备抗击粗差的能力,为此,统计学家作了大量的工作,才逐渐形成稳健估计。稳健估计一般分为M、L、R 估计。M 估计是极大似然估计,由于易于实现,应用也较广。在计算中,针对灰色建模数据序列的残差为ε_i($i = 1,2,\cdots,n$) ,M 估计的函数可取为$\rho(\varepsilon)$,M 估计的准则为$\sum_{i=1}^{n}\rho(\varepsilon_i) = \min$; M 估计中的ρ是任意选取的函数,其稳定性与ρ的选则有关。选取的ρ不同,会出现不同的M 估计,稳定性也不一样。在M 估计的许多方法中,计算较为简单的是选权迭代法。选权迭代法的计算过程为:①根据灰色微分方程列立误差方程$\varepsilon = Y - Ba$,并令各观测序列值的权函数值为1 ,即$p_1(\varepsilon_1) = p_2(\varepsilon_2) = \cdots = p_n(\varepsilon_n) = 1$。②利用最小二乘法答解出a 和ε的第1 次估计值, $a^{(1)} = (B^TB)^{-1}B^TY$; $\varepsilon^{(1)} = Y - Ba^{(1)}$。③由ε来确定各个观测序列值的权函数$p_i(\varepsilon_i)$,再重新建立误差方程,答解出$a^{(2)}$和$\varepsilon^{(2)}$,与迭代计算一样,直到前后2 次计算的差值符合要求为止。④最后的结果为$a_k = (B^TP^{(K-1)}(\varepsilon)B)^{-1}B^TP^{(K-1)}(\varepsilon)Y$; $\varepsilon = Y - Ba_k$。对于上述的计算过程,其中随着函数ρ的选取不同,就对应着不同的权函数形式,并且在迭代的过程中ρ值随着计算数值的变化不断的变化,从而达到根据数据的变化来调整数据在其计算中的权重。

现在几种常用的选权迭代法包括Huber 法、一次范数最

作者简介　李双柱(1967 -) ,男, 山西平朔人,高级工程师,从事测绘工程技术及研究。

收稿日期　2008-10-23

小法(L^1 估计)、p 范最小法(L^p 估计)、Hampel 法、IGG(周文江法)、验后方差估计法(李德仁法)。

采用一次范数最小法(L^1 估计),此计算法中确定权函数的方法。*ρ函数为* $p(\varepsilon) = |\varepsilon|$;相应权函数为 $p(\varepsilon) = 1/|\varepsilon|$;为了避免 $\varepsilon = 0$ 时出现权的计算问题,改为 $p(\varepsilon) = 1/(|\varepsilon| + \delta)$。式中,*$\delta$相对于 ε是一个微小量。*

一次范数最小法(L^1 估计)采用残差的绝对值的倒数来确定权重,从而充分把数据序列精度通过残差体现出来,既然该方法可以根据数据的精度来确定权重,完全可利用稳健估计,即一次范数最小法(L^1 估计)来进行灰色建模计算。因为考虑到数据的精度问题[8-11],就会使得数据精度高的把数据本身的数据信息体现在灰色建模中。另外,一次范数最小法(L^1 估计)还具有抗击粗差的能力,因此对少数据序列建模的灰色模型也是一个非常好的方法,能提高数据信息的利用率。

3 数据来源及实例分析

3.1 数据来源 试验地设在湖北省建始县国营长岭岗林场内,该林场位于北亚热带高山气候区,30°40′N,109°30′E,海拔1 600~1 900 m,年均温度为10 ℃左右,相对温度85%左右,全年降雨量为1 500~2 000 mm,土壤肥沃且呈微酸性,总面积为1 023 hm^2。自20世纪50年代末引种以来,现已成为当地主要的速生丰产树种产区之一,对其研究显然很有必要。为此,选择位于不同海拔、坡向、林龄等因子的林分,在全场范围内设面积为625 m^2 的标准地43块,在每块标准地内进行每木检尺,按径阶测取5~10株树高。在树高曲线上查出林分条件平均高后,根据平均胸径和条件平均高,选定平均标准木,以2年为龄阶,整理出与龄阶相对应的胸径、树高、材积,数据见表1。

表1 日本落叶松平均胸径、平均树高、平均材积原始序列

Table 1 The original series data on DBH, mean tree height and average volume of Larix Kaempferi

龄阶 Stage of age	平均胸径 DBH//cm	平均树高 H//m	平均材积 V//m^3	龄阶 Stage of age	平均胸径 DBH//cm	平均树高 H//m	平均材积 V//m^3
10	9.6	9.00	0.047 0	24	17.2	19.45	0.219
12	11.5	11.10	0.055 4	26	17.8	20.40	0.256
14	12.7	12.49	0.078 0	28	19.0	21.55	0.293
16	13.8	13.50	0.097 2	30	19.8	22.20	0.311
18	14.4	15.40	0.118 0	32	20.3	24.46	0.373
20	15.7	17.01	0.141 0	34	21.8	26.15	0.458
22	16.1	18.30	0.167 0				

3.2 实例分析一 选取表1中的胸径数据序列中9.6、11.5、12.7、13.8、14.4、15.7、16.1、17.2 作为研究对象 $X^{(0)}$。对 $X^{(0)}$ 作1-AGO,得 $X^{(1)} = (x^{(1)}(1), x^{(1)}(2), \cdots, x^{(1)}(13)) = (9.6, 21.1, 33.8, 47.6, 62, 77.7, 93.8, 111)$。

对 $X^{(1)}$ 作紧邻均值生成,得

$Z^{(1)} = (z^{(1)}(2), z^{(1)}(3), \cdots, z^{(1)}(13))$

$= (15.35, 27.45, 40.70, 54.80, 69.85, 85.75, 102.40)$;利用最小二乘估计和稳健估计分别对上述数据建立灰色模型,得到时间响应式,最小二乘估计为 $x^{(1)}(k+1) = 182.9e^{0.063k} - 173.3$;稳健估计为 $x^{(1)}(k+1) = 193.8e^{0.06k} - 184.2$;用上述2个公式计算出 $x^{(1)}$ 的模拟值和 $x^{(0)}$ 的还原值,2种计算方法的比较见表2。

表2 模型拟合比较

Table 2 Modd fitting comparison

序号 No	原始值 Original value	最小二乘估计 Least squares estimation		稳健估计 Robust estimation	
		模拟值 Simulated value	残差 Residuals	模拟值 Simulated value	残差 Residuals
2	11.5	11.89	-0.39	11.98	-0.48
3	12.7	12.67	0.03	12.73	-0.03
4	13.8	13.49	0.31	13.51	0.29
5	14.4	14.37	0.03	14.35	0.05
6	15.7	15.30	0.40	15.23	0.47
7	16.1	16.30	-0.20	16.18	-0.08

由表2可知,2种计算方法得到的模拟值都与原值接近,残差相对较小。就总体而言,稳健估计除个别数据的残差大于最小二乘估计外,绝大部分都小于最小二乘估计。出现这种情况是最小二乘估计均衡误差的原因。

3.3 实例分析二 对表2中的原始数据作人为附加误差。将4号数据变为15.8,7号数据变为17.1,2种估计方法的计算比较见表3。

表3 抗击粗差的比较

Table 3 Combat gross error comparison

序号 No	原始值 Original value	最小二乘估计 Least squares estimation		稳健估计 Robust estimation	
		模拟值 Simulated value	残差 Residuals	模拟值 Simulated value	残差 Residuals
2	11.5	12.39	-0.89	11.85	-0.35
3	12.7	13.16	-0.46	12.68	0.02
4	13.8	13.98	-0.18	13.58	0.22
5	14.4	14.84	-0.44	14.53	-0.13
6	15.7	15.75	-0.05	15.55	0.15
7	16.1	16.73	-0.63	16.65	-0.55

由表3可知,稳健估计的拟合明显好于最小二乘估计。最小二乘估计的残差序列明显带有规律性,证明人为加入误差干扰项后,拟合曲线发生移动。

3.4 实例分析三 某林场1982~1991年的病虫害发病率是25%、39%、42%、33%、21%、17%、18%、16%、17%、15%,对其利用最小二乘估计进行灰色预测,建立 GM(1,1) 模型(表4)。

表4 最小二乘估计计算

Table 4 Least squares estimation calculation

序号 No	原始数据 Original value	模拟值 Simulated value	残差 Residuals	相对误差//% Relative error
2	39	41.45	-2.45	6.3
3	42	35.54	6.46	15.3
4	33	30.41	2.59	7.8
5	21	26.02	-5.02	23.9
6	17	22.26	-5.26	30.9
7	18	19.05	-1.05	5.8
8	16	16.29	-0.29	1.8
9	17	13.94	3.06	18.0
10	15	11.93	3.07	20.5

(下转第10页)

解、接触酶、H_2S 试验结果均为阴性。通过菌落形态、个体形态及生理生化反应结果,初步鉴定 S_5 菌株为乳杆菌。1974 年,北京大学生物系酸浆研究小组研究认为生产绿豆淀粉用酸浆的主要作用菌是乳酸链球菌[2];2006 年,刘文菊等对生产绿豆淀粉用的酸浆中的乳链球菌进行了研究,结果表明绿豆淀粉酸浆中的乳链球菌具有凝集作用[3]。试验通过在实验室自然发酵生产的甘薯酸浆进行研究,利用 MRS 培养基对酸浆中的乳酸菌进行分离,并通过对分离出的 11 株乳酸菌进行淀粉沉淀能力测定,得到一株沉淀淀粉能力较好的菌株,经初步鉴定为乳杆菌。这说明乳杆菌菌体本身或由其产生的代谢产物对淀粉也具有絮凝作用。

表 1 分离出的乳酸菌菌株对淀粉的沉淀能力

Table 1 The precipitation capacity of the isdated *Lactobacillus* strains on starch

菌株 Strain	pH 值 pH value	沉淀时间 Precipitation time	菌株 Strain	pH 值 pH value	沉淀时间 Precipitation time
S_1	6.23	53.32	S_8	3.76	30.83
S_2	5.76	35.52	S_9	5.52	76.22
S_3	5.68	31.62	S_{10}	5.39	44.31
S_4	6.13	66.35	S_{11}	5.41	36.17
S_5	5.01	11.03	S_{12}	6.12	73.29
S_6	4.12	23.16	S_{13}	5.34	57.88
S_7	5.23	56.11			

从发酵液的乳酸菌总数、pH 值来看,乳酸菌总数偏高的发酵液,pH 值一般较低,分离效果也较好[6]。这说明分离效果与乳酸菌总数及酸浆的 pH 值有关,但究竟是乳酸菌菌体本身还是 pH 值或菌体的其他代谢产物对淀粉沉淀起主要作用,试验还得不到结论,有待进一步研究。

3 结论

(1) 从自然发酵的甘薯酸浆中分离出 13 株乳酸菌,通过发酵试验筛选出一株分离效果良好的乳酸菌,并通过该菌的群体形态特征、个体形态特征观察及生理生化试验,初步鉴定该乳酸菌为乳杆菌。

(2) 试验表明乳杆菌对淀粉也具有一定的絮凝作用。

参考文献

[1] 郑玮,沈群. 化学处理对酸浆中一株淀粉凝集菌和乳酸乳球菌 As1.9 沉淀淀粉能力的影响[J]. 食品工业科技,2007(10):123-126.

[2] 北京粉丝厂,北京大学生物系酸浆研究小组. 酸浆为什么能沉淀淀粉[J]. 北京大学学报,1974(1):57-63.

[3] 杜连起,刘绍军. 酸浆作用菌对甘薯淀粉沉淀效果的研究[J]. 现代商贸工业,1998(8):39-41.

[4] 曹宗巽,卢光莹,宋云,等. 乳酸链球菌凝集淀粉粒机理的进一步研究[J]. 食品科学,1980,20(3):271-275.

[5] 刘文菊,沈群,刘杰. 酸浆法生产淀粉机理研究初探[J]. 食品科学,2006,27(1):79-82.

[6] 秦礼康,江萍,张倩,等. 微生物发酵酸浆用于马铃薯淀粉生产工艺研究[J]. 贵州农业科学,1997,25(5):42-44.

(上接第 6 页)

由表 4 可知,个别几个数据的拟合相对误差较大,达不到建模要求,可采取利用残差修正的方法。为了解能否利用稳健估计来实现该数据序列的灰色建模,采取稳健估计和灰色新陈代谢理论相结合处理,得到较好的拟合精度(表 5)。

表 5 稳健估计计算

Table 5 Robust estimation calculation

序号 No	原始数据 Original value	模拟值 Simulated value	残差 Residuals	相对误差 // % Relative error
2	21	18.81	2.19	10.4
3	17	17.99	-0.99	5.8
4	18	17.23	0.77	4.3
5	16	16.48	-0.48	3.0
6	17	15.77	1.23	7.2
7	15	15.09	-0.09	0.6

由表 5 可知,拟合精度完全满足建模要求。证明在灰色建模过程中,可以利用稳健估计来代替最小二乘估计,因为稳健估计从数据的精度出发来确定权重,充分考虑了数据的精度问题,因此拟合的精度高一些。

4 讨论

(1) 通过对稳健估计在林业灰色建模中应用的探讨,从理论和实例上说明了稳健估计在灰色建模中完全可以替代利用最小二乘法建立灰色模型,建模精度完全满足灰色建模要求。利用稳健估计来建立模型,可以替代残差修复模型的建立,减少了建模程序。

(2) 利用稳健估计来对数据列建立灰色模型,可以剔出数据序列中的粗差、提高建模的精度。

(3) 由于灰色建模是利用少数据序列信息来建立模型,还有很多问题需要解决。稳健估计考虑了数据的精度问题,但并不能解决所有的问题。应该从数据序列的实际出发来考虑问题,才可以找到解决的办法。

(4) 数据列中存在的缺失,是一个非常实际而且经常要遇到的问题。既然稳健估计可以抗击粗差,那么可以把缺失的数据按照粗差作为研究对象,这应该是一个解决缺失数据的办法。

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] 朱坚民,王中宇. 测量数据粗大误差的非统计判别[J]. 华中理工大学学报,2004(4):17-19.

[3] WANG Z Y,XIA X T,ZHU J M. Research development of the grey error throry and the application in the dynamic measurement[J]. Fifth International Symposium on Instrumentation and Contrd Technology,SPIE,2003,5253:447-451.

[4] 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.

[5] 刘思峰. 灰色系统理论的产生与发展[J]. 南京航空航天大学学报,2004(2):267-272.

[6] 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.

[7] 陶本藻. 稳健估计应用问题[J]. 地矿测绘,2000(1):1-3.

[8] 梁长秀,冯仲科,郎南军,等. 森林资源调查数据的稳健估计及分析[J]. 北京林业大学学报,2001(6):10-12.

[9] 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.

[10] 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.

[11] 姜岩,刘文生,范学理. 稳健估计在岩移参数辨识中的应用[J]. 阜新矿业学院学报,1996(7):278-282.