基于分形插值的我国旱灾数据分析研究

王 萍[1],倪丽萍[1],倪 洋[2]
(1.合肥工业大学 管理学院,安徽 合肥 230009;
2.山东财经大学 金融学院,山东 济南 250002)

摘 要:由于旱灾的发生是自然、社会等多种影响因素共同作用的结果,旱灾数据往往呈现出复杂性和非线性。文中根据旱灾数据的特点,以1974年至2004年我国旱灾成灾面积为例,运用分形理论对其进行分析。首先,使用R/S法对旱灾成灾面积时间序列的分形特征进行判别,结果显示该时间序列具有良好的分形特征;然后利用分形插值方法实现了对历史数据的插值拟合;最后根据建立的分形插值预测模型,对2005年的旱灾成灾面积进行预测。实验结果表明,拟合和预测的结果与实际情况都较为接近。因此,利用分形插值方法来分析旱灾数据是合理的。

关键词:分形理论;R/S分析;曲线拟合;预测
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2015)08-0199-04
doi:10.3969/j.issn.1673-629X.2015.08.042

Research on Analysis of Chinese Drought Data Based on Fractal Interpolation

WANG Ping[1], NI Li-ping[1], NI Yang[2]
(1. School of Management, Hefei University of Technology, Hefei 230009, China;
2. Institute of Finance, Shandong University of Finance and Economics, Jinan 250002, China)

Abstract: The occurrence of drought is the result of combined action which caused by natural, social and other factors, so the drought data often presents complexity and nonlinear. According to the characteristics of drought data, taking the data of Chinese drought disaster area that occurs in 1974 to 2004 as an example, use the fractal theory to analyze. Firstly, apply the R/S analysis method to determine the fractal characteristics of drought disaster area time series, the results show that this time series has good fractal characteristics. Then, use fractal interpolation method to achieve the historical data interpolation fitting. Finally, apply the fractal interpolation prediction model which is established to forecast the drought disaster area in 2005. Experimental results show that the fitting and predictive results are relatively close to the actual situation. Therefore, using the fractal interpolation method to analyze the drought data is reasonable.

Key words: fractal theory; R/S analysis; curve fitting; forecast

0 引 言

旱灾是我国最常见的自然灾害之一,具有持续时间长、波及范围广、影响力度大等特征[1]。每年的夏季是旱灾爆发的高峰期,长时间的高温和降雨量的不足很容易引发水资源短缺、农作物减产。旱灾往往会给受灾地区带来巨大的经济损失,例如,2013年7月~8月,安徽、浙江等7个南方省市因旱灾就导致了农作物受灾面积达12 000多万亩,其中绝收1 700多万亩,直接经济损失共几百亿元。因此,研究历史旱灾数据、挖掘数据背后的信息是十分必要的。如果可以正确地掌握旱灾的演变规律,提前制定相应的抗旱方案,那么就能在一定程度上降低灾害的影响,这对保护生态环境、维护社会稳定具有重要意义。

长期以来,国内外专家学者们使用了各种不同的分析方法,试图从不同的角度揭示旱灾的特性。例如:周振民等[2]以历年旱灾情况为基础,将BP神经网络与Z指数法相结合建立了旱灾预测模型,有效实现了灾情预测;杜灵通[3]以山东省为例,对山东近10年来的干旱历程进行了定量监测,运用多种空间数据时序分析方法,研究了山东干旱的时空演化特征;Mehdi Rezaeian-Zadeh等[4]利用优化的多层感知器网络,实

收稿日期:2014-09-24 **修回日期**:2014-12-30 **网络出版时间**:2015-07-21
基金项目:国家自然科学青年基金项目(71301041)
作者简介:王 萍(1989-),女,硕士研究生,研究方向为数据挖掘、人工智能。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150721.1448.048.html

现了对干旱指数 SPI 的预测,比较了不同时间尺度下的预测效果;Akyuz Dilek-Eren 等[5]通过分析葡萄牙 67 年的 SPI 数据,运用马尔可夫链揭示了旱灾的随机性;Ganguli Poulomi[6]将支持向量机与 Copula 函数相结合建立了干旱预测模型,提高了干旱预测的能力。

尽管这些方法已经取得了一些成果,但旱灾是一个典型的复杂、不规则系统,旱灾数据往往呈现出非线性和无条理性,所以传统的数据分析方法不能很好地对它做深入研究。文中根据旱灾数据的特点,试图用分形理论进行分析。分形方法能够从看似杂乱无章的数据中找出数据变化的规律,是解决非线性问题的有效方法之一[7]。文中以我国几十年的旱灾受灾数据为研究对象,在判定该时间序列数据具有分形特征的基础上,引入了分形插值方法,实现了对旱灾数据的拟合和预测。

1 R/S 分析

1965 年,英国著名的水文学家 Hurst 在大量实证研究的基础上提出了 R/S(Rescaled Range Analysis)分析法[8-9]。目前,该方法已被广泛运用于时间序列分形特征的判别。在 R/S 分析法中,区分时间序列是随机的还是非随机的主要取决于一个数量指标—Hurst 指数(H)。若统计量 $H=0.5$,表明该时间序列是完全随机的;若 H 在 0 ~ 0.5 之间,意味着该时间序列具有反持续性;若 H 在 0.5 ~ 1 之间,标志该时间序列具有长期的正持续性。

对于一时间序列 $\{X(i) \mid i=1,2,\cdots,n\}$,可以定义序列的均值为:

$$X_\tau = \frac{1}{\tau}\sum_{i=1}^{\tau} X(i), \tau = 1,2,\cdots,n \tag{1}$$

累计离差:

$$X(i,\tau) = \sum_{k=1}^{i}[X(k) - X_\tau], 1 \leqslant i \leqslant \tau \tag{2}$$

极差:

$$R(\tau) = \max X(i,\tau) - \min X(i,\tau) \tag{3}$$

标准差:

$$S(\tau) = \sqrt{\frac{1}{\tau}\sum_{i=1}^{\tau}[X(i) - X_\tau]^2} \tag{4}$$

研究发现,如果 $R(\tau)$ 与 $S(\tau)$ 的比值满足 $R/S \propto \tau^H$ 关系,则说明该时间序列存在 Hurst 现象[10]。为了计算方便,可将式子两边同时取对数,则 $\log(R/S)$ 与 $\log(\tau)$ 所构成图形的斜率即为 H 的值。

2 分形插值方法

2.1 分形插值函数

对于传统的数学插值方法,如 Lagrange 插值、Newton 插值、Hermite 插值等,插值函数一般由多项式、有理函数等线性组合表示,而分形插值函数则是通过迭代函数系(IFS)来实现[11-12]。对于一个二维的数据集 $\{(x_i,y_i), i=0,1,\cdots,N\}$,可构造 IFS$\{R^2: W_i, i=1,2,\cdots,N\}$,其中每个函数 W_i 都是仿射变换,具体形式可表示为:

$$W_i\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}a_i & 0\\c_i & d_i\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} + \begin{bmatrix}e_i\\f_i\end{bmatrix} \tag{5}$$

式中,a_i,c_i,d_i,e_i,f_i 为五个未知常数,当确定它们的取值后,就可以得到某一特定的仿射变换形式。由式(5)可知,一次仿射变换的过程可以简述为 $(x,y) \rightarrow (a_ix + e_i, c_ix + d_iy + f_i)$,因此,求函数 W_i 也就是求这几个常数的取值。

在构造 IFS 时,每个 W_i 必须要满足如下两个条件:

$$W_i\begin{bmatrix}x_0\\y_0\end{bmatrix} = \begin{bmatrix}x_{i-1}\\y_{i-1}\end{bmatrix} \tag{6}$$

$$W_i\begin{bmatrix}x_N\\y_N\end{bmatrix} = \begin{bmatrix}x_i\\y_i\end{bmatrix} \tag{7}$$

将式(6)、(7)的左侧按照式(5)的形式展开,可得到关于 a_i,c_i,d_i,e_i,f_i 的四个方程,即

$$\begin{cases} a_ix_0 + e_i = x_{i-1} \\ a_ix_N + e_i = x_i \\ c_ix_0 + d_iy_0 + f_i = y_{i-1} \\ c_ix_N + d_iy_N + f_i = y_i \end{cases} \tag{8}$$

一般将 d_i 作为自由变量,则通过式(8)可以解得其他四个参数的值。

$$\begin{cases} a_i = \dfrac{x_i - x_{i-1}}{x_N - x_0} \\ c_i = \dfrac{y_i - y_{i-1}}{x_N - x_0} - d_i\dfrac{y_N - y_0}{x_N - x_0} \\ e_i = \dfrac{x_Nx_{i-1} - x_0x_i}{x_N - x_0} \\ f_i = \dfrac{x_Ny_{i-1} - x_0y_i}{x_N - x_0} - d_i\dfrac{x_Ny_0 - y_Nx_0}{x_N - x_0} \end{cases} \tag{9}$$

2.2 分形插值迭代过程

若有一简单图形 L,迭代函数系中的函数 $W_i(i=1,2,\cdots,N)$ 都已知。在第一次迭代时,每个 W_i 都对 L 进行压缩变换,则可以得到 N 个与 L 自相似的小图形,分别记为 $W_1(L),W_2(L),\cdots,W_N(L)$,将这 N 个结果拼贴在一起就形成了第一次迭代的输出图形 L_1,即 $L_1 = W_1(L) \cup W_2(L) \cup \cdots \cup W_N(L)$。

在第二次迭代时,以新产生的 L_1 为基础图形,每个 W_i 都对 L_1 进行压缩变换,同样,可以得到 N 个与 L_1 自相似的小图形,输出图形 $L_2 = W_1(L_1) \cup W_2(L_1) \cup \cdots \cup W_N(L_1)$。

每次将上一轮的输出作为下一轮的初始图形,不断重复上面的迭代过程就可以形成一个复杂的分形图形。

2.3　基于分形插值的预测模型

如果一时间序列数据具有分形特征,由分形的自相似性和标度不变性可知[13],将该时间序列向外延拓后,延拓部分应当与区间内部分具有相同的分形特征[14]。根据这个思想,可以构建基于分形插值的预测模型,具体步骤如下:

(1)获取原始时间序列数据 $\{(x_i,y_i) \mid i=0,1,\cdots,n\}$。

(2)数据归一化处理,$y_i^* = (y_i - y_{\min})/(y_{\max} - y_{\min})$,标准化结果记为 $A=\{(x_i,y_i^*) \mid i=0,1,\cdots,n\}$。

(3)从 A 中挑选进行分形插值的插值点,构成集合 $B=\{(X_i,Y_i) \mid i=0,1,\cdots,N\}$。

(4)根据需要预测的时间确定预测点横坐标 X_{N+1},纵坐标 Y_{N+1} 赋初值为 0,将预测点 (X_{N+1},Y_{N+1}) 作为新插值点加入集合 B。

(5)按公式(8)计算插值点集合 B 的迭代函数系,再根据分形插值迭代原理迭代 k 次,得到第 k 次迭代产生的点集 C。

(6)从 C 中挑选横坐标最接近原始数据 $x_i(i=0,1,\cdots,n)$ 的 $n+1$ 个点,记为 $M=\{(a_i,b_i) \mid i=0,1,\cdots,n\}$。

(7)计算 $S=\dfrac{1}{n+1}\sum_{i=0}^{n} \mid b_i - y_i{}^* \mid$。

(8)选取合适的步长 δ,从 0 到 1 逐渐改变集合 B 中 Y_{N+1} 的大小,不断重复步骤(5)~(7),找到 S 值最小时所对应的 Y_{N+1},再通过步骤(2)中的公式还原,即可得到最终的预测值。

3　实验过程及结果

3.1　旱灾数据的分形特征分析

数据选取:时序数据来自中国统计网,提取 1974~2004 年全国的旱灾成灾面积(千公顷),将该数据作为原始时序数据。

通过对旱灾成灾面积时间序列进行 R/S 分析,可以得到图 1 所示的结果。

图中,lg$t$ 与 lg(R/S) 之间的线性关系比较明显,计算得到的离散点基本上分布在一条直线上。运用最小二乘法对离散点进行拟合,可以得到图中的拟合直线,拟合直线的斜率即为 Hurst 指数的估计值,经计算得到 $H=0.668\ 3$。H 在 0.5~1 之间,表明了该时间序列不是随机序列,数据之间存在着一定的时间相关性,分形特征明显。根据分形维数的计算公式 $D=2-H$,可以求得该时间序列的分形维数 D 为 1.331 7。

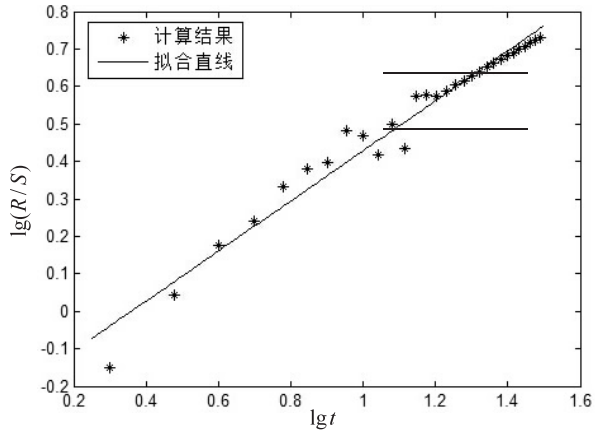

图 1　旱灾成灾面积时间序列 R/S 分析

3.2　分形插值在旱灾数据拟合中的应用

根据已知,1974~2004 年共有 31 年的历史数据,现运用分形插值方法对这些数据进行插值拟合。为了方便表示,将 1974 年至 2004 年分别用序号 1,2,…,31 代替,选取其中的 10 年数据作为已知插值点数据(见表 1)。

表 1　插值点数据

时间/年	序号	面积/千公顷
1974	1	2 740
1977	4	7 010
1979	6	9 316
1980	7	14 174
1984	11	7 015
1989	16	15 262
1990	17	7 805
1998	25	5 060
2000	27	26 784
2004	31	8 482

将插值点集合代入公式(5)~(9)进行运算,可以得到一个包含 9 个仿射变换的迭代函数系 IFS,具体的 IFS 码如表 2 所示。利用该迭代函数系对已知数据进行迭代,迭代后可以得到相应的分形插值拟合结果。

表 2　迭代函数系的 IFS 码

IFS 码	a_i	c_i	d_i	e_i	f_i
1	0.100 0	104.05	0.200 0	0.900 0	2 087.9
2	0.066 7	0.3067	0.400 0	3.933 3	5 913.7
3	0.033 3	152.36	0.050 0	5.966 7	9 026.6
4	0.133 3	−244.37	0.030 0	6.866 7	14 336
5	0.166 7	255.76	0.100 0	10.833	6 485.2
6	0.033 3	−267.71	0.100 0	15.967	15 256
7	0.266 7	−254.19	0.850 0	16.733	5 730.2
8	0.066 7	704.99	0.100 0	24.933	4 081.0
9	0.133 3	−611.98	0.010 0	26.867	27 369

从分形插值拟合结果中提取出时间轴分别为 1,2,…,31 共 31 个坐标点及其对应的函数值。将分形插值得到的 31 个拟合函数值与实际值进行比较,比较

结果如图 2 所示。

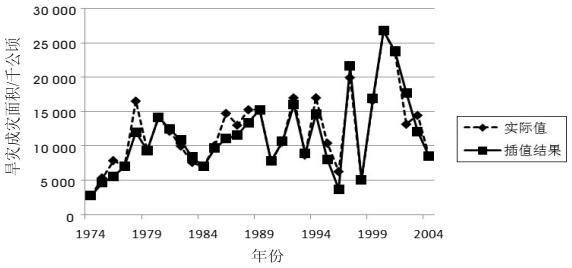

图 2 插值结果与实际数据对比图

从图中可以明显看出,两条数值曲线的变化形式和数值大小都极为相似,实际值曲线中的几次数据波动在分形插值结果中几乎都有所体现。

分形插值方法能取得如此好的拟合效果,不仅仅是因为分形插值方法本身是研究不规则图形的有力工具,同时,也反映了旱灾成灾面积时间序列确实具有良好的分形特性,序列中数据的变化具有较高的自相似性。因此,运用分形插值方法来研究该序列数据是非常合理的。

3.3 分形插值在旱灾数据预测中的应用

现根据 1974 ~ 2004 年的历史数据,利用文中建立的分形插值预测模型,对 2005 年的旱灾成灾面积(N)进行预测。该模型为了简化计算,将所有年份的旱灾成灾面积进行归一化处理,预测值 N 的取值范围也缩小到了[0,1]之间。图 3 展示了 N 取不同值(0.2,0.4,0.6,0.8)时的插值结果。

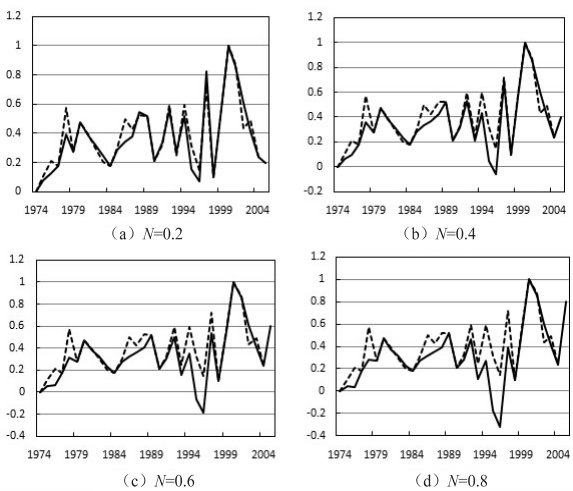

图 3 预测点取不同值时的插值结果

通过比较可以看出,随着 N 取值的不同,迭代插值的结果也不同。图中, N = 0.2 时插值结果最接近原始数据, N = 0.8 时插值结果偏离实际数据最多。通过逐步改变 N 的大小,可以求得最终的预测值,将结果还原后可以得到 2005 年的旱灾成灾面积约为 8 030千公顷,而 2005 年实际旱灾成灾面积为 8 479 千公顷,虽然计算结果有一定的误差,但总体还是较为接近的,并且符合 2005 年数值下降的趋势。

4 结束语

文中以 1974 ~ 2004 年我国旱灾成灾面积时间序列为研究对象,在验证了该时间序列具有分形特征的基础上,应用分形插值方法对序列中的数据进行插值拟合,同时,利用文中建立的分形插值预测模型,对 2005 年的旱灾成灾面积进行了预测。实验结果表明,利用分形插值方法来拟合和预测旱灾成灾面积序列数据是可行的,并且,拟合和预测的结果与实际情况也较为接近。

虽然分形插值方法在处理某些复杂的序列数据时具有一定的优势,但是,该方法在应用过程中还存在一些有待解决的问题。例如,目前计算迭代函数系中 d_i 的方法还不够科学;插值点选取还没有统一的规则等。这些问题也是未来在分形插值方面的研究课题。

参考文献:

[1] 孙仲益,张继权,严登华,等. 基于 GIS 的安徽省旱灾风险空间演变研究[J]. 东北师大学报:自然科学报,2012,44(4):133-137.

[2] 周振民,谢滨帆. BP 神经网络在郑州市旱灾预测中的应用及防灾减灾对策[J]. 中国农村水利水电,2011(12):97-100.

[3] 杜灵通. 基于多源空间信息的干旱监测模型构建及其应用研究[D]. 南京:南京大学,2013.

[4] Rezaeian-Zadeh M,Tabari H. MLP-based drought forecasting in different climatic regions[J]. Theoretical and Applied Climatology,2012,109(3):407-414.

[5] Akyuz D E, Bayazit M, Onoz B. Markov chain models for hydrological drought characteristics[J]. Journal of Hydrometeorology,2012,13(1):298-309.

[6] Ganguli P,Reddy M J. Ensemble prediction of regional droughts using climate inputs and the SVM-copula approach[J]. Hydrological Processes,2014,28(19):4989-5009.

[7] Mandelbrot B. The fractal geometry of nature[M]. New York: W. h. Freeman,1982.

[8] 李 蔚. 江苏海岛气候序列分数维分析及分形理论应用[D]. 南京:南京信息工程大学,2009.

[9] 朱良燕,毛军军,苗 强,等. 合肥市降水变化趋势分形特征分析与预测[J]. 计算机技术与发展,2009,19(9):17-20.

[10] 尹晓惠,王式功. 我国北方沙尘暴与强沙尘暴过程的分形特征及趋势预测[J]. 中国沙漠,2007,27(1):130-136.

[11] 樊昭磊. 关于分形插值函数若干分析性质的研究[D]. 南京:南京财经大学,2011.

[12] 李水根,吴纪桃. 分形与小波[M]. 北京:科学出版社,2002.

[13] Hutchinson J E. Fractal and self-similarity[J]. Indiana Univ. Math J,1981,30(5):713-747.

[14] 刘 鑫. 分形插值法在中国证券市场指数分析中运用[D]. 南京:南京理工大学,2009.

基于分形插值的我国旱灾数据分析研究

作者： 王萍， 倪丽萍， 倪洋， WANG Ping， NI Li-ping， NI Yang

作者单位： 王萍,倪丽萍,WANG Ping,NI Li-ping(合肥工业大学 管理学院,安徽 合肥,230009)， 倪洋,NI Yang(山东财经大学 金融学院,山东 济南,250002)

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2015(8)