第 59 卷 第 12 期 2010 年 12 月
1000-3290/2010/59(12)/8986-08

物 理 学 报
ACTA PHYSICA SINICA

Vol. 59, No. 12, December, 2010

去趋势波动分析方法中不重叠等长度子区间长度的确定*

侯 威[1)2)†] 章大全[1)3)] 杨 萍[4)] 杨 杰[1)3)]

1)(国家气候中心,北京 100081)

2)(中国科学院大气物理研究所东亚区域气候环境重点实验室,北京 100029)

3)(兰州大学大气科学学院,兰州 730000)

4)(中国气象局北京城市气象研究所,北京 100089)

(2010 年 1 月 12 日收到;2010 年 6 月 29 日收到修改稿)

针对去趋势波动分析方法中参数不重叠等长度子区间长度 s 的选取,基于信息论的基本原理,提出使用符号分析方法对原始数据进行符号编码,并使用不同的方式对符号序列进行分段、计算互信息函数. 细致描述了不同分段方式对原始混沌序列的信息编码能力,以此判断所采用的分段方式能否真实有效地还原原始序列所包含的全部信息. 给出了确定最优分段个数或各分段长度的具体方式,确定了不重叠等长度子区间长度 s 的选取算法,以及判断所研究序列是否适用于去趋势波动分析方法,避免了以往参数 s 选取中随机性和主观性给计算结果带来的错误信息. 进一步将该方法应用于实际温度资料,计算并分析中国 1961—2000 年逐日平均温度的去趋势波动分析指数分布状况.

关键词: 去趋势波动分析, 参数选择, 区间, 最优选取

PACC: 9260X

1. 引 言

为了可靠地检测时间序列中的持续性特征,就必须严格区分外在趋势和序列自身固有的长期波动特征. 外在趋势往往会使时间序列表现出缓慢的单调变化或单摆波动特征,从而得出错误的序列持续性定量特征. 20 世纪 90 年代一些学者提出了一种全新的研究时间序列波动长程相关性的标度指数计算方法,即去趋势波动分析(detrended fluctuation analysis,简记为 DFA)方法[1]. DFA 方法的优势在于它消除了局部趋势,能准确地观察到时间序列本身所具有的统计行为特征,避免了将时间序列的短程相关、非平稳性虚假地检测为长程相关性.

DFA 方法被证明是检测非平稳时间序列长程相关性的最重要、最可靠的工具之一,已成功地应用于脱氧核糖核酸序列、心率序列、云层结构、地质学和金融学等方面[2—12]. 可是,关于 DFA 方法中如何确定不重叠等长度子区间的长度 s,目前还没有一种统一方法,大都依据经验取值,具有一定的随机性和主观性. 当参数 s 取不同值时,得到的 DFA 指数也存在较大差别,甚至得到错误的信息. 因此,迫切需要一种基于序列本身且客观实际的方法来统一参数 s 的选取,以确保计算所得长程相关指数的合理性和正确性.

2. DFA 方法

DFA 方法是基于随机过程理论和混沌动力学新发展的一种分析方法. 从动力学角度看,这种方法中变换后的序列仍保留着原序列的痕迹,与原序列保持着相同的持久性(或反持久性);同时,变换可较好地"滤除"序列本身演化的趋势,剩下的离差序列主要为波动成分. 因此在分析非平稳时间序列时,采用 DFA 方法可以避免对相关性的错误判断.

* 国家自然科学基金(批准号:40905034, 40775048)和国家科技支撑计划(批号:2007BAC29B01, 2009BAC51B04)资助的课题.

† E-mail: hou_w@sohu.com

在 DFA 方法的基础上，Kantelhardta 等[13]于 2002 年又进一步提出非平稳有限序列的多重分形去趋势波动分析（multifractal detrended fluctuation analysis，简记为 MF-DFA）方法. MF-DFA 方法不仅可以检测长程相关性、确定其标度不变性，即分形结构特征，还能判定序列是否具有多重分形属性并确定其多分形特征.

对于长度为 N 的序列 $\{x_k, k=1,2,\cdots,N\}$，下面给出 DFA 方法的具体分析过程.

首先，建立一新序列

$$y(i) = \sum_{k=1}^{N} [x_k - \langle x \rangle] \qquad (i = 1,2,\cdots,N), \tag{1}$$

式中 $\langle x \rangle$ 为原序列 $\{x_k\}$ 的均值. 将新序列 $y(i)$ 划分成长度为 s 的不重叠等长度子区间，长度为 N 的序列共被分为 $L = \mathrm{Int}(N/s)$（$\mathrm{Int}(\cdot)$ 表示对括号内数值取整）个子区间. 因序列长度 N 不一定被子区间长度 s 整除，为保证原序列信息不丢失，可以从序列末端开始反向前再划分一次，这样可得到共 $2N_s$ 个子区间. 对每个子区间 $v(v=1,2,\cdots,2N_s)$ 的数据进行多项式回归拟合，得到局部趋势函数 $y_v(i)$，$y_v(i)$ 可以是一阶，二阶或更高阶的多项式，分别记为 DFA1，DFA2 等. 消除各子区间内趋势，计算其方差均值，

$$F^2(v,s) = \frac{1}{s}\sum_{i=1}^{s} \{y[(v-1)s+i] - y_v(i)\}^2 \qquad (i = 1,2,\cdots,N_s), \tag{2}$$

$$F^2(v,s) = \frac{1}{s}\sum_{i=1}^{s} \{y[N-(v-N_s)s+i] - y_v(i)\}^2 \qquad (i = N_s+1, N_s+2,\cdots,2N_s). \tag{3}$$

然后，确定全序列的 q 阶波动函数

$$F_q(s) = \left\{\frac{1}{2N_s}\sum_{v=1}^{2N_s} [F^2(v,s)]^{q/2}\right\}^{1/q}, \tag{4}$$

式中 q 可以取为任何非零实数. $q=0$ 时，(4) 式变为

$$F_0(s) = \exp\left\{\frac{1}{4N_s}\sum_{v=1}^{2N_s} \ln[F^2(v,s)]\right\}. \tag{5}$$

最后，通过分析双对数坐标图 $F_q(s)$-s 关系，可以确定波动函数的标度指数 h_q，即存在幂律关系

$$F_q(s) \propto s^{h_q}. \tag{6}$$

自相关函数一般用于静态数据的分析，而 DFA 对不稳定信号和噪声信号不敏感. 不同的尺度指数值反映了不同的时间过程，h_q 与自相关函数有着密切的关系[14,15].

在本文的计算中，对每个子区间 $v(v=1,2,\cdots,2N_s)$ 的数据进行二阶多项式回归拟合，并取 $q=2$. 对于平稳序列而言，h_q 就是赫斯特指数 H. 对于非平稳序列，当 $h_q=0.5$ 时，该序列为一独立过程；当 $0.5<h_q\leqslant 1$ 时，该序列存在长程相关性；当 $h_q<0.5$ 时，该序列存在负长程相关[16—18].

3. 不同参数 s 对长程相关指数的影响

目前 DFA 方法在自然科学和社会科学等领域已得到了广泛的应用[2—12]，但 DFA 方法中如何确定不重叠等长度子区间的长度 s，尚无统一方法，大都依据经验定义为 $q+2\leqslant s\leqslant N/4$（$q$ 为 MF-DFA 方法的阶数，N 为所研究序列的长度），受随机性和主观性影响较大. 并且对于同一序列，由于取不同的参数时所得到的 DFA 指数（即序列的长程相关指数）会有很大的差别. 本文首先分析了在理想时间序列中不同的参数 s 对 DFA 方法计算结果的影响，指出现有方法存在的问题，并进一步以信息理论为基础，采用联合信息熵的方法来确定参数 s 的选取.

所用理想序列为一满足高斯分布的随机序列，样本量为 10000，根据经验公式 $q+2\leqslant s\leqslant N/4$，参数 s 最大值 $s_{\max}$ 的取值范围为 [100,1000]，最小值 $s_{\min}$ 的取值范围为 [−10,10]，讨论不同 s 值对其 DFA 指数的影响. 图 1 为参数 s 取不同值时得到的长程相关指数. 从图 1 可以看出，当参数 s 取不同值时，DFA 指数 $-6<h_q\leqslant 8$，彼此之间差异很大且分布在

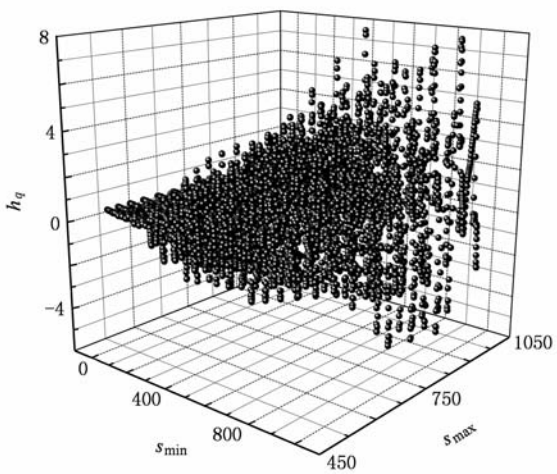

图 1　参数 s 取不同值时得到的长程相关指数 h_q

如此之大的值域内也是完全不合理的,无法真正得到系统本质的长程相关性;并且 $s_{\max}$ 越大,$s_{\min}$ 越小,计算得到的 DFA 指数的差异就越显著.

(1)式中的 $y(i)$ 实际代表了原始序列 $\{x_k\}$ 的演变趋势,因而当 $s_{\max}$ 和 $s_{\min}$ 的取值相差很大时,会导致由(4)式得到的 $F_q(s)$ 以及最终得到的 DFA 指数也有很大差异. 由于理想序列是一随机序列,所以其 DFA 指数 $h_q=0.5$.

从以上分析可知,即使对于同一个序列或系统而言,也可能仅仅因为在 $q+2\leqslant s\leqslant N/4$ 区间中取了不同的 s 值,而对该序列或系统长程相关性的判断得到完全不一样甚至相反的研究结论,阻碍对研究对象的物理本质属性的认知.

4. 用互信息确定参数 s 的算法

4.1. 算法设计

下面将基于信息论基本原理[19—21],采用符号分析方法计算互信息函数,确定 DFA 方法中参数 s 的选取,并在此基础上通过若干典型的数值计算来验证方法的有效性. 对于一个含有 N 个元素的变量 $x(N)$,记 $p_x(x_i)$ 为变量 $x(N)$ 处于状态 $x(i)$ 的概率,则变量 $x(N)$ 的信息熵定义为[22]

$$H(x)=-\sum_{i=1}^{N}p_x(x_i)\mathrm{lb}[p_x(x_i)], \tag{7}$$

另一变量 $y(N)$ 相对于变量 $x(N)$ 的条件熵定义为

$$\begin{aligned}H(y/x)&=-\sum_{i,j}^{N}p_{x,y}(x_i,y_j)\mathrm{lb}[p_{x,y}(x_i,y_j)/p_x(x_i)]\\&=H(x,y)-H(x).\end{aligned} \tag{8}$$

这里 $p_{x,y}(x_i,y_j)$ 是变量 $x(N)$ 和 $y(N)$ 分别处于状态 x_i,y_j 时的联合概率,$H(x,y)$ 是变量 $x(N)$ 和 $y(N)$ 的联合熵,有

$$H(x,y)=-\sum_{i,j}^{N}p_{x,y}(x_i,y_j)\mathrm{lb}[p_{x,y}(x_i,y_j)]. \tag{9}$$

变量 $x(N)$ 和 $y(N)$ 之间的互信息定义为

$$\begin{aligned}I(y;x)&=H(y)-H(y/x)\\&=H(x)+H(y)-H(x,y)\\&=I(x;y).\end{aligned} \tag{10}$$

对于互信息的计算,Fraser 等[22]给出了一种等概率划分空间格子的方法,这种方法原理上不易掌握. 另一种是等间距划分空间格子的方法[23,24],这种方法虽然简便,但如何确定划分的间距又带有一定的随意性. 这两种方法的计算都很繁琐,不便于应用. 本文使用符号分析的方法来计算互信息[25—30],该方法计算简便、精确,是一种较好的计算互信息的方法.

给定一个由 m 个符号组成的符号集 $\{u_0,u_1,\cdots,u_{m-1}\}$ 和一个 $m+1$ 个临界点 $\{x_{c0},x_{c1},\cdots,x_{cm}\}$ 组成的集合,若 $x_{ck}\leqslant x_j\leqslant x_{ck+1}$,则 $Q(j)=u_k$. 一次对原始序列 $x(N)$ 进行粗粒化,将序列 $x(N)$ 转化为一个符号序列 $Q(N)$,进一步将得到的符号序列 $Q(N)$ 依次分成 R 段,每段长度 $L=\mathrm{Int}(N/R)$,各段记为 $L_1,L_2,\cdots,L_R$,并通过下式来标记和辨识这些短序列[25—30]:

$$T_R(L,i)=\sum_{k=1}^{L}m^{L-k}Q(k+i-1), \tag{11}$$

式中 i 表示短序列在符号序列 $Q(N)$ 中是从第 i 个符号开始. 将符号 u_k 用对应的整数 k 来代替,这样每一个短序列都可以使用整数集 $\{0,1,2,\cdots,m^L-1\}$ 中的某一个整数来进行唯一地标记和辨别,原始时间序列 $x(N)$ 可进一步转换成由这些短序列为元素依次组成的符号序列 $T(R)$. 用 p_{T_R} 表示特定短序列 $T_R(L,i)$ 出现的概率,它可以用短序列 $T_R(L,i)$ 在序列 $T(R)$ 中出现的次数除以所有短序列的总数来计算. 时间序列 $x(N)$ 包含在符号序列 $T(R)$ 中的信息熵可以表示为

$$H(T)=-\frac{1}{L}\sum_{p_{T_R}}p_{T_R}\mathrm{lb}(p_{T_R}). \tag{12}$$

由上述规则定义的符号语言对混沌序列的信息编码能力与分割临界点 x_{cm} 的数目及取值有关,一般可以通过使熵 $H(T)$ 取到最大值的方式来得到最优临界点的数目和取值.

下面以另一种方式对符号序列 $Q(N)$ 进行分段[31]. 将符号序列 $Q(N)$ 分成 $r=L=\mathrm{Int}(N/R)$ 段,每段长度为 $l=R$,记为 $l_1,l_2,\cdots,l_r$,各段包含的符号标记为 $u_1(r),u_2(r),\cdots,u_r(r)$,包含在各个小段 l_r 中的信息熵表示为

$$H(l_r)=-\frac{1}{l}\sum_{p_u}p_u\mathrm{lb}(p_u), \tag{13}$$

式中 p_u 为各个小段 l_r 中不同符号 $u_r(r)$ 出现的概率.

对于序列 $T(R)$ 和每个短序列 l_r,其联合熵可以表示为

$$H(T,l_r)=-\frac{1}{R}\sum_{T_R}\sum_{l_r}p(T_R,l_r)\mathrm{lb}[p(T_R,l_r)]. \tag{14}$$

这里的 $p(T_R,l_r)$ 是变量 $T(R)$ 处于状态 $T_R(L,i)$ 而变量 $u_r(r)$ 处于状态 u_m 的联合概率，它可以用变量 $T(R)$ 处于状态 $T_R(L,i)$ 而变量 $u_r(r)$ 处于状态 u_m 的联合序列数除以符号序列 $T(R)$ 和 $u_r(r)$ 的联合序列的总数来计算. 因此，时间序列 $T(R)$ 与每个短序列 l_r 之间的互信息可表示为

$$\begin{aligned} I(R) &= H(T) + H(l_r) - H(T,l_r) \\ &= -\frac{1}{L}\sum_{p_{T_R}} p_{T_R}\mathrm{lb}(p_{T_R}) - \frac{1}{l}\sum_{p_u} p_u\mathrm{lb}(p_u) \\ &\quad + \frac{1}{R}\sum_{T_R}\sum_{l_r} p(T_R,l_r)\mathrm{lb}[p(T_R,l_r)]. \end{aligned} \tag{15}$$

由上述过程可知，$H(T)$ 是原始序列 $x(N)$ 所包含的信息，序列 $T(R)$ 与每个短序列 l_r 之间的互信息 $I(r)$ 实际代表了各个短序列 l_r 所包含的关于原始序列 $x(N)$ 的信息，则 $T(R)$ 与所有短序列 l_r 的互信息之和的平均 $\overline{I(R)} = \sum_{i=1}^{r} I(i)/r$ 就可以认为是对原始序列 $x(N)$ 进行分割之后得到的一系列子序列平均所具有的关于 $x(N)$ 的信息量.

图 2(a) 给出了对样本量为 10000 满足高斯分布的随机序列 $x(N)$ 取不同的 R 值时，包含在符号序列 $T(R)$ 中的 $x(N)$ 的信息熵 $H(T)$、平均每个短序列 l_r 所包含的信息熵 $\overline{H(l_r)}$ 以及序列 $T(R)$ 和各个短序列 l_r 的平均联合熵 $\overline{H(T,l_r)}$ 的变化曲线，R 取区间[10,1000]中的连续整数.

结合本文提出的算法可知，因 $L = \mathrm{Int}(N/R)$，故当 R 值较小时，L 值较大，此时原始序列 $x(N)$ 仅仅被分割成几段，$H(T)$ 仅仅包含少量关于 $x(N)$ 的信息并可能失真，符号序列 $T(R)$ 并不能反映 $x(N)$ 的演化特征. 同时因 $r = L = \mathrm{Int}(N/R)$，故 r 值较大，l 值较小，原始序列 $x(N)$ 被分割成若干小段 l_r，各 l_r 的 $\overline{H(l_r)}$ 同样只包含很少量的信息，因而平均联合熵 $\overline{H(T,l_r)}$ 也同样较小. 随着 $x(N)$ 的分割段数(即 R 值)的不断增大，包含在符号序列 $T(R)$ 中关于 $x(N)$ 的信息也在增多，符号序列 $T(R)$ 可以逐步反映出 $x(N)$ 的演化特征，R 值的不断增大导致 r 值减小，原始序列 $x(N)$ 被分割的段数越来越少，l 值变大，各 l_r 计算得到的 $\overline{H(l_r)}$ 中所包含的信息增多，$\overline{H(T,l_r)}$ 也随之增加. 较之 $H(T)$ 和 $\overline{H(T,l_r)}$，$\overline{H(l_r)}$ 随 R 值的增大要缓慢得多，这是因为当 l 值足够大时，各小段的 $\overline{H(l_r)}$ 已包含足够多的信息，再增大 l 值所带来的信息增量微乎其微. 对于 $H(T)$ 而言，R 越大，对 $x(N)$ 进行分割的次数越多，直至趋向于 $R=1$(此时 $T(R)$ 反映 $x(N)$ 演化特征的能力最强)，

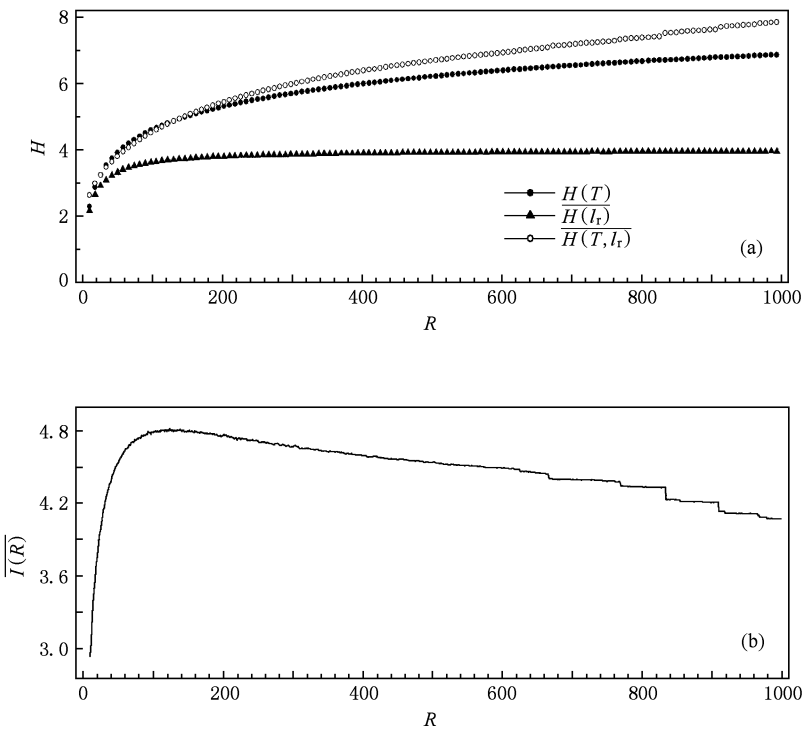

图 2　高斯分布的随机序列的信息熵 $H(T)$，$\overline{H(T,l_r)}$，$\overline{H(l_r)}$ 和平均信息量 $\overline{I(R)}$ 随 R 的变化　(a)随机序列的信息熵随 R 的变化，(b) 平均信息量随 R 的变化

自然会带来更多关于原始序列 $x(N)$ 的信息,也使得 $H(T,l_r)$ 保持较大的增加趋势. 此时 $H(T,l_r)$ 中的大部分贡献来自于 $H(T)$.

图 2(b)为各 l_r 中所具有的关于 $x(N)$ 的平均信息量 $\overline{I(R)}$ 随 R 的变化. 从图 2(b)中可以看出,当 R 值较小时,$\overline{I(R)}$ 也较小,且随着 R 值的增大而增大. 由上述分析可知,这是因为当 R 值较小时,$H(T)$ 仅仅包含少量关于 $x(N)$ 的信息,各 l_r 的 $\overline{H(l_r)}$ 同样只包含很少量的信息,所以包含在各 l_r 中的关于 $x(N)$ 的信息量 $\overline{I(R)}$ 也很小. 随着 R 值的增大,$H(T)$ 和 $\overline{H(l_r)}$ 包含的信息也越来越多,各个子序列 l_r 的演变反映了 $x(N)$ 演化特征的能力也在不断地增强,自然 $\overline{I(R)}$ 也随之变大,最终在 $R=R'$ 时 $\overline{I(R)}$ 达到最大(R' 为 $\overline{I(R)}$ 达到最大值时所对应的 R 值). 此时各 l_r 中包含的关于 $x(N)$ 的信息最多,其反映 $x(N)$ 演化特征的能力达到顶峰,能最大限度地捕捉到 $x(N)$ 的演化特征. 随着 R 值的进一步增大,$\overline{I(R)}$ 值却开始逐渐减小,这是因为 R 值越大,$\overline{H(l_r)}$ 的增加越小,$H(T)$ 与 $\overline{H(T,l_r)}$ 的增加却相对较大,并且在 $R>R'$ 以后 $H(T)$ 的增率开始小于 $\overline{H(T,l_r)}$ 的增率,导致在 $R>R'$ 时 $\overline{I(R)}$ 表现出下降的趋势. 说明当 R 值过大时,各 $H(l_r)$ 中都含有一部分冗余信息,使得 $\overline{H(T,l_r)}$ 中具有虚假的综合信息,反而使 $\overline{I(R)}$ 变小,各 l_r 反映 $x(N)$ 演化特征的能力受到削弱.

由以上分析可知,使用符号分析方法并对符号序列 $Q(N)$ 使用不同的方式进行分段、计算互信息,可以细致描述各个分段对原始混沌序列的信息编码能力,以判断所采用的分段方式是否合适,即能否真实有效地还原原始序列所包含的全部信息. 下面将给出确定最优分段个数或各分段长度的具体思路及检验.

综上所述,$R=R'$ 时各个 l_r 中关于 $x(N)$ 的信息量 $\overline{I(R)}$ 最多,反映 $x(N)$ 演化特征的能力最强,此时对原始序列 $x(N)$ 的分割是最佳的,因此将 DFA 方法中参数 s 的上限取为 $s\leqslant R'$,图 2(b)中 $R'=124$,据此得到理想序列 $x(N)$ 的 $s\leqslant 124$. 对于 s 的下限,因为 R 太小时 $\overline{I(R)}$ 也很小且有失真的可能,因此取 s 的下限为 $s\geqslant R'\times 0.9$,保证了此时的 $\overline{I(R)}$ 包含足够的真实信息且处于上升趋势之中,计算得到 $s\geqslant 37$.

4.2. 算法检验

对理想高斯分布随机序列,取 $37\leqslant s\leqslant 124$ 计算得到的 DFA 指数为 $h_q=0.489$,接近理论值 0.5,如图 3 所示. 上述算法是取 $\overline{I(R)}$ 最大值的 90% 时 $\overline{I(R)}$ 所对应的 R 值作为参数 s 的下限,这样的取值虽然有一定的主观因素,但 s 下限的取值并不固定. 图 3 同时也给出了将 s 的下限分别调整为 $s-5$ 和 $s+5$,同时保持 s 的上限值不变时得到的 DFA 指数. $32\leqslant s\leqslant 124$ 和 $42\leqslant s\leqslant 124$ 计算得到的 DFA 指数分别为 $h_q=0.491$ 和 $h_q=0.495$,结果也同样接近理论值 0.5.

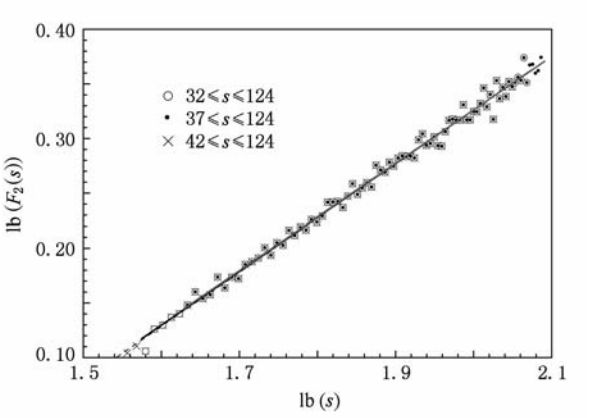

图 3 高斯分布理想随机序列 s 值下限取不同值时得到的 DFA 指数

按上述算法进一步计算了相同长度的指数分布和平均分布理想随机序列,得到参数 s 的选取范围分别为 $57\leqslant s\leqslant 176$ 和 $22\leqslant s\leqslant 73$,其 DFA 指数分别为 $h_q=0.494$ 和 $h_q=0.506$. 将两者 s 值的下限分别调整为 $s-5$ 和 $s+5$,并保持 s 的上限值不变计算其 DFA 指数,指数分布理想随机序列的结果分别为 $h_q=0.508$ 和 $h_q=0.505$,均匀分布理想随机序列的结果为 $h_q=0.491$ 和 $h_q=0.498$,均接近理论值 0.5. 这说明了本文算法的有效性.

由于理想随机序列不可能是真正的随机序列,因而会产生计算结果与理论值之间的偏差. 虽然 s 的下限值在 $[-5,+5]$ 范围内波动,但计算得到 DFA 指数之间只有很微小的差别,并且其值也都非常符合理论值 0.5. 这说明该算法关于参数 s 的下限值取值方法是有效的,当 s 的下限值在理想值附近波动时不会影响 DFA 指数的计算结果.

已有研究表明,Henon 映射具有典型的短程相关性,而 Lorenz 模型则具有典型的长程相关性[32,33]. 从表 1 可以看出,当序列长度较长时,不同 s 值计算得到的 DFA 指数值彼此之间差异很小,并且非常接近原始序列的 DFA 指数值,说明尽管序列的长度发生了变化,但对于同一系统,使用本算法

得到的结果是稳定的. 当序列长度变为 1000 个数据点时,不同 s 值对应的 DFA 指数值之间差异较大,与理论值相比差异也较大.

时间序列的长程相关性反映了系统本身的演化特征,是系统自身固有属性. 这种属性包含在反映系统演变行为的时间序列之中,因此并不随着表征系统演化的时间序列长度的变化而变化. 如果这一序列没有包含系统的足够信息,即序列本身不能反映系统的演变特征,此时计算得到的 DFA 指数自然也不能真实反映系统的长程相关性.

表 1　不同模型生成的理想序列的 DFA 指数

序列长度	Henon 模型		Lorenz 模型		指数分布		均匀分布		高斯分布	
	s	指数	s	指数	s	指数	s	指数	s	指数
10000	46—175	0.282	65—214	0.942	17—73	0.508	52—196	0.483	32—124	0.495
	51—175	0.285	70—214	0.930	22—73	0.506	57—196	0.494	37—124	0.491
	56—175	0.283	75—214	0.938	27—73	0.505	62—196	0.491	42—124	0.489
5000	47—175	0.279	62—214	0.921	24—99	0.499	52—198	0.465	35—124	0.529
	52—175	0.295	67—214	0.977	29—99	0.510	57—198	0.496	40—124	0.474
	57—175	0.307	72—214	0.846	34—99	0.501	62—198	0.468	45—124	0.523
1000	43—109	0.232	58—138	0.145	25—110	0.395	51—148	0.666	40—137	0.598
	48—109	0.349	63—138	3.164	30—110	0.572	56—148	1.580	45—137	0.530
	53—109	0.472	68—138	0.238	35—110	0.759	61—148	0.333	50—137	−0.121

图 4 给出了 Lorenz 混沌系统的 X 分量序列长度为 1000 个数据点,s 值下限取不同值时得到的 DFA 指数拟合情况. 对于单个 s 取值区间,在 lb(s)-lb($F_2(s)$)图中的数据点与拟合直线偏差较大. 从 DFA 分析方法的计算步骤看,说明当序列长度较短时各段的演变差异巨大,即使最优的分段结果只包含了极少量的信息,导致不同 s 值取值范围得到的 $F_2(s)$ 也具有很大差异,也就是意味着对这一序列进行的分段去趋势处理是不合适的,该序列只含有控制系统演变的片面信息.

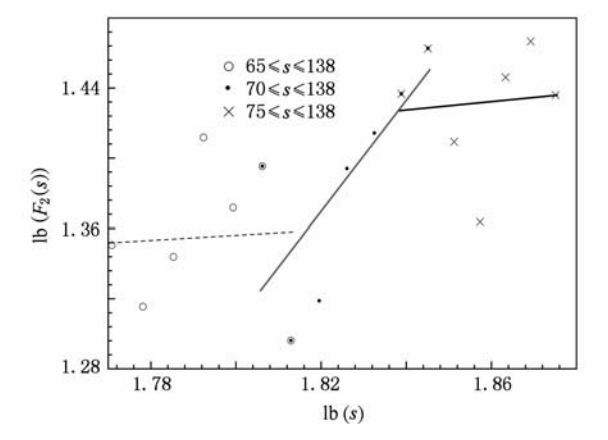

图 4　Lorenz 混沌系统的 X 分量 s 值下限取不同值时得到的 DFA 指数拟合情况

对于长度为 10000 和 5000 的序列而言,当 s 值下限取不同值时,在 lb(s)-lb($F_2(s)$)图中(图略)的点则比较集中,重合现象很明显. 这说明当序列长度较长时,最优的分段方式使各段的演变有较多的相似之处,在不同的 s 取值范围内得到的 $F_2(s)$ 也只有较少的数据点存在差异,各段均包含了较多的信息并出现了共同拥有的信息,序列包含了控制系统演变的足够信息,表明对该长度的序列 DFA 方法的处理是恰当而合适的. 同时,还进一步验证了其他长度的序列,限于篇幅,本文没有一一给出,但得到的结果完全支持上述结论.

4.3. 实际应用

将 DFA 方法应用于实际资料,分析其长程相关性,计算结果如图 5 所示. 在计算过程中采用本文提出的方法来确定参数不重叠等长度子区间长度 s 的选取. 所用资料为中国气象局公布的中国 165 个国际交换站 1961—2000 年无缺测的逐日实测平均温度资料,所有资料均是直接使用,没有经过二次处理.

从图 5 可以看出,DFA 指数 h_q 整体呈东高西低的分布状态且其值均大于 0.5. 这表明实际温度序列中的各个值之间不是独立的,具有长程相关性,是一个具有持久性的增强时间序列,即在 t 时刻以前存在上升(或下降)趋势隐含着 t 时刻以后总体上也存在上升(或下降)趋势,其趋势增强行为取决于 $h_q>0.5$ 的程度. DFA 指数 h_q 值在江淮东部及江南

东部地区最大,该地区温度序列的长程相关性也最高,趋势增强行为最明显,说明该地区气候系统的记忆性较好,即存在较高的可预测性[15];华北、内蒙古及广大中部地区的 h_q 值为次高;西藏、新疆和东北地区的 h_q 值较小;西南地区东南部、华南地区南部和川西高原地区的 h_q 值最小,也就是长程相关性最低,趋势增强行为也最不显著,气候系统的记忆性及可预测性相对较差.

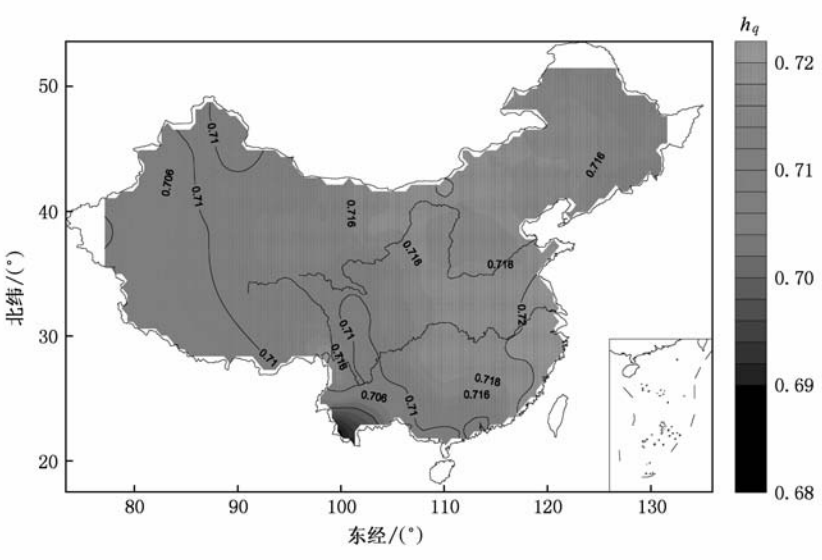

图 5 中国 1961—2000 年逐日平均温度 DFA 指数 h_q 的分布

5. 结 论

基于信息论基本原理,采用符号分析的方法计算互信息函数,确定 DFA 方法中参数 s 的选取,通过几种典型的数值试验验证了方法的可行性. 这一算法完全基于数据本身,因而具有较强的客观性,且该算法对于序列长度不敏感.

当序列长度较短时,即使最优的分段结果只包含了极少量的信息,导致在不同的 s 取值范围内得到的 $F_2(s)$ 也有很大差异,意味着对这一序列进行的分段去趋势分析是不合适的,该序列只含有控制系统演变的片面信息. 当参数 s 值的下限在一定范围内(本文取[-5, +5])波动时,得到的 DFA 指数差异不明显,说明序列包含了足够多的系统演化信息,可以使用本文提出的算法来确定系统的 DFA 指数;反之,如果得到的 DFA 指数有明显差异,则该序列只包含系统演化的部分信息,由此得到的关于系统的 DFA 指数是虚假的,该序列不适用于 DFA 方法. 进一步将该方法应用于实际温度资料,计算并分析了中国 1961—2000 年逐日平均温度的 DFA 指数分布状况.

衷心感谢丑纪范院士为本文工作提出宝贵意见.

[1] Peng C K, Buldyrev S V, Havlin S 1994 *Phys. Rev.* E **49** 1685
[2] Lux T, Marehesi M 1999 *Nature* **397** 498
[3] Mantegna R N, Stanley H E 1995 *Nature* **376** 46
[4] Liu F, Shan X M, Ren Y, Zhang J, Ma Z X 2004 *Acta Phys. Sin.* **53** 550 (in Chinese)[刘 锋、山秀明、任 勇、张军、马正新 2004 物理学报 **53** 550]
[5] Sealas E 1998 *Physica* A **253** 394
[6] Janosi I M, Janeesko B, Kondor D 1999 *Physica* A **269** 111
[7] Feng G L, Wang Q G, Hou W, Gong Z Q, Zhi R 2009 *Acta Phys. Sin.* **58** 2853 (in Chinese)[封国林、王启光、侯 威、龚志强、支 蓉 2009 物理学报 **58** 2853]
[8] Stanley H E, Amaral L A N, Canning D 1999 *Physica* A **269** 156
[9] Stanley H E, Afanasyevr V, Anlaral L A N 1996 *Physica* A **224** 302
[10] Wang Q G, Zhi R, Zhang Z P 2008 *Acta Phys. Sin.* **57** 5343 (in Chinese)[王启光、支 蓉、张增平 2008 物理学报 **57** 5343]
[11] He W P, Wu Q, Zhang W, Wang Q G, Zhang Y 2009 *Acta Phys. Sin.* **58** 2862 (in Chinese)[何文平、吴 琼、张 文、

王启光、张　勇 2009 物理学报 **58** 2862]
[12] Govindan R B, Vjushin D, Brenner S 2001 *Physica* A **294** 239
[13] Kantelhardta J W, Zschiegner S A, Bunde E K 2002 *Physica* A **316** 87
[14] Bernaola G P 2001 *Phys. Rev. Lett.* **87** 168
[15] Wang Q G, Hou W, Zheng Z H, Gao R 2009 *Acta Phys. Sin.* **58** 6640 (in Chinese)[王启光、侯　威、郑志海、高　荣 2009 物理学报 **58** 6640]
[16] Panlov A N, Sosnovtseva O V, Ziganshin A R 2002 *Physica* A **316** 233
[17] Lee J M, Kin D J, Kim I Y 2002 *Comp. Bio. Med.* **32** 37
[18] Ott E 1993 *Chaos in Dynamical Systems* (Cambridge: Cambridge University Press) pp305—333
[19] Yang X L, Xu W 2008 *Chin. Phys.* B **17** 2004
[20] Zhang D Z 2007 *Acta Phys. Sin.* **56** 3152 (in Chinese)[张佃中 2007 物理学报 **56** 3152]
[21] Xiao F H, Yan G R, Han Y H 2005 *Acta Phys. Sin.* **54** 550 (in Chinese)[肖方红、阎桂荣、韩宇航 2005 物理学报 **54** 550]
[22] Fraser A M, Swinney H L 1986 *Phys. Rev.* A **33** 1134
[23] Yang Z A, Wang G R, Chen S G 1995 *Chin. J. Comp. Phys.* **12** 442 (in Chinese)[杨志安、王光瑞、陈式刚 1995 计算物理 **12** 442]
[24] Nichols J M, Nichols J D 2001 *Math. Biosci.* **171** 21
[25] Rechester A B, White R B 1991 *Phys. Lett.* A **156** 419
[26] Rechester A B, White R B 1997 *Phys. Rev. Lett.* **78** 54
[27] Lehrman M, Rechester A B 2001 *Phys. Rev. Lett.* **87** 164
[28] Liu Z H, Chen S G 1997 *Phys. Rev.* E **56** 7297
[29] Azad R K, Rao J S, Ramaswamy R 2002 *Chaos Solitions Fract.* **14** 633
[30] Xiao F H, Yan G R, Han Y H 2004 *Acta Phys. Sin.* **53** 2877 (in Chinese)[肖方红、阎桂荣、韩宇航 2004 物理学报 **53** 2877]
[31] Zheng Z H, Ren H L, Huang J P 2009 *Acta Phys. Sin.* **58** 7359 (in Chinese)[郑志海、任宏利、黄建平 2009 物理学报 **58** 7359]
[32] Lehrman M, Rechester A B 2001 *Phys. Rev. Lett.* **87** 164501
[33] Liu Z H, Chen S G 1997 *Phys. Rev.* E **56** 7297

A valid method to compute the segment size in detrended fluctuation analysis*

Hou Wei [1)2)†]　Zhang Da-Quan [1)3)]　Yang Ping[4)]　Yang Jie[1)3)]

1) (*National Climate Center*, *Beijing* 100081, *China*)

2) (*Key Laboratory of Regional Climate Environment Research for Temperate East Asia*, *Institute of Atmospheric Physics*, *Chinese Academy of Sciences*, *Beijing* 100029, *China*)

3) (*College of Atmospheric Sciences*, *Lanzhou University*, *Lanzhou* 730000, *China*)

4) (*Institute of Urban Meteorology of Beijing*, *China Meteorological Administration*, *Beijing* 100089, *China*)

(Received 12 January 2010; revised manuscript received 29 June 2010)

Abstract

We develop a method to compute the segment size in the detrended fluctuation analysis (DFA), which is based on the basic concept of the information theory, and verify the method effectiveness by numerical experiment. This method is freed from the problem of subjectivity in the former process to choose the segment size which usually leads to false result. We Change the length of sequence with dynamics being the same, the results remain stable. The results indicate that when the length of sequence is too short, even the optimal selection of segment size is not enough for the portrait of the overall dynamic system, thus the DFA cannot be used in this circumstance. The method we developed in this paper can enhance the reliability of DFA results by judging whether the sequences analyzed meet the requirements of DFA. We also obtain the DFA index from 1961 to 2000 of China through DFA method and analyze its spatial characteristics of distribution.

Keywords: detrended fluctuation analysis, parameter selection, segment, optimal selection

PACC: 9260X

* Project supported by the National Natural Science Foundation of China (Grant Nos. 40905034, 40775048) and the State Key Program of Science and Technology of China (Grant Nos. 2007BAC29B01, 2009BAC51B04).

† E-mail: hou_w@sohu.com