海表面温时间序列的相关性及复杂性研究

于文静,余 洁,徐凌宇
(上海大学 计算机工程与科学学院,上海 200444)

摘 要: 海洋表面温度是海洋热力、动力过程以及海气相互作用的综合结果,是海洋环境一个重要的参数。为了获得海洋表面温度时空变化的复杂性行为并揭示其内在动力性机制,根据一个改进的样本熵的方法来分析海表面温时间序列的复杂性;同时使用去趋势波动分析方法研究海表面温时间序列的长记忆性。实验结果表明,高纬度地区的海表面温变化的复杂度低于低纬度地区。因为高纬度地区海表面温度的季节性更加明显,序列的规则性更强。相对来说季节性因素对低纬度地区海表面温度的影响不大。不管是长时间序列还是短时间序列高纬度地区的海表面温都具有长记忆性,且序列的不平稳性很强。但是短时间序列的低纬度地区的海表面温的相关性呈现出多种情况,序列足够长时,序列也表现出长记忆性。

关键词: 海表面温; 复杂度; 二维熵; 相关性; 去趋势波动分析
中图分类号: TP31 **文献标识码:** A **文章编号:** 1673-629X(2019)02-0181-04
doi: 10.3969/j.issn.1673-629X.2019.02.038

Research on Correlation and Complexity of Sea Surface Temperature Time Series

YU Wen-jing, YU Jie, XU Ling-yu
(School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract: Sea surface temperature, as an important parameter of marine environment, is the comprehensive result of ocean thermal and dynamic processes as well as air sea interaction. In order to obtain the complexity of the temporal and spatial variation of ocean surface temperature and reveal its intrinsic dynamic mechanism, the complexity of the sea surface temperature time series is analyzed based on an improved sample entropy method, and the long-term memory of that is studied by the method of detrended fluctuation analysis. The experiment shows that the complexity of sea surface temperature change in high latitudes is lower than that in low latitudes. Because the seasonality of sea surface temperature is more obvious in high latitudes, and the sequence is more regular. Generally speaking, seasonal factors have little effect on sea surface temperature in low latitudes. For long time series or short time series, the sea surface temperature in high latitudes is long-term memory and the sequence is not stable. However, the correlation of sea surface temperature at low latitudes in short time series shows many situations. When the sequence is long enough, the sequence also shows long-term memory.

Key words: sea surface temperature; complexity; TD_entropy; correlation; detrended fluctuation analysis

0 引 言

海洋系统是一个动态的复杂系统,在演化过程中,海洋系统动力学特征往往呈现出复杂性和非线性。海洋表面温度(sea surface temperature,SST)是一个重要的海洋环境参数,几乎所有的海洋过程,特别是海洋动力过程都直接或间接地与温度有关。SST 时间序列具有明显的季节性,长记忆性行为是复杂系统内在机制的综合表现,是探索时空耦合系统的深层动力机制的有效工具。因此对 SST 时间序列的复杂性及相关性进行研究,对揭示海洋的特性和机理具有非常重要的意义[1-3]。

样本熵[4]是一种衡量时间序列复杂度的方法,广泛应用在生理信号、脑电信号、振动信号等复杂度分析中[5-6]。但是样本熵在衡量时间序列复杂度时,熵值的大小有时和系统复杂度的增加没有关系[7],因此文中提出了一种样本熵的改进算法——二维熵。相比于

收稿日期: 2018-04-09 **修回日期:** 2018-08-02 **网络出版时间:** 2018-11-15
基金项目: 科技部重点研发计划(2016YFC1401902)
作者简介: 于文静(1990-),女,硕士研究生,研究方向为不确定信息挖掘; 余 洁,副教授,研究方向为网络个性化搜索、用户兴趣建模、基于语义的网络交互计算等; 徐凌宇,教授,研究方向为基于 Web 的远程软件服务技术、网络多源信息融合技术、大规模数据挖掘、数字地球技术。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20181115.1046.020.html

样本熵,二维熵在计算向量的相似性时,不仅仅考虑向量之间的模式相似性,还考虑了向量之间的时间距离对相似性的影响。同时,二维熵在最后计算熵值的模型中,考虑了序列中向量的自相似程度对时间序列复杂性的影响。因此二维熵在计算时间序列的复杂性时更加合理。

相关性的分析最早是由英国水文学家 H. E. Hurst[8] 于 1954 年提出的,通过重标极差(R/S)分析法构建 Hurst 指数来度量时间序列的长程相关性。1992 年,C.K.Peng 等在《Nature》上提出的基于随机游走理论的标准标度分析法[9],计算出波动指数可以刻画时间序列的长程相关性和短程相关性。但这两种方法都是基于平稳时间序列建立的,而现实世界中大多数的时间序列都是非平稳的。对于非平稳的时间序列,C.K.Peng 等于 1994 年在研究 DNA 序列时提出了去趋势波动分析法(detrended fluctuation analysis, DFA)[10],能够有效减小序列局部趋势导致的非平稳性。DFA 分析方法已成功应用在金融市场、情感心电信号、气温变化、降雨演变中等[11-15]。

1 方 法

1.1 二维熵方法

二维熵的定义和样本熵类似,二维熵参数用 N, m,r,k 表示。其中,N 为序列长度,m 为嵌入维数,r 为相似容限,k 为时间衰减系数,具体算法如下:

设原始时间序列 $U(i)$ 为由 N 个点构成的序列,根据预先设定的嵌入维数 m 将原始时间序列重构成一组 m 维向量,每个向量代表从第 i 个点开始连续的 m 个 u 的值:

$$X_i^m = \{u(i), u(i+1), \cdots, u(i+m-1)\}, \quad i = 1,2,\cdots,N-m+1 \tag{1}$$

定义向量 X_i^m 和 X_j^m 间的距离为 $d_{i,j}^m$,用欧氏距离计算为两个向量对应元素差值最大的一个,即:

$$d_{i,j}^m = \max|u(i+k) - u(j+k)|, k \in [0, m-1], \quad i \neq j \tag{2}$$

根据给定的相似容限 r,衰减系数 k,以及 X_i^m 和 X_j^m 间的距离 $d_{i,j}^m$,得到两个向量的相似性为:

$$\mu_{i,j}^m = \begin{cases} 1 \times e^{-k \times |i-j|}, d_{i,j}^m \leqslant r \\ 0, d_{i,j}^m > r \end{cases} \tag{3}$$

统计每个向量 X_i^m 和其他向量 X_j^m 之间相似可能性的概率和,并求出与匹配的向量总数 $N-m-1$ 的比值,记为 $B_i^m(r,k)$,代表序列中 X_i^m 与其他向量相似可能性的平均概率。

$$B_i^m(r,k) = \frac{\sum_{i=1}^{N-m} u_{i,j}^m}{N-m-1} \tag{4}$$

然后计算每个向量 X_i^m 和其他向量 X_j^m 相似可能性的平均概率的和,并除以序列中 m 维向量的总数,得到 m 维向量的自相似的概率,记为 $B^m(r,k)$。

$$B^m(r,k) = \frac{\sum_{i=1}^{N-m} B_i^m}{N-m} \tag{5}$$

将维数 m 增加 1,重复上述步骤,得到 $B^{m+1}(r,k)$。

二维熵在计算时间序列的复杂度时不仅考虑新信息的产生率还考虑向量自相似程度,故二维熵的计算模型如下:

$$\text{TD\_entropy}(m,r,k) = \lim_{N\to\infty} -\ln\left[\frac{B^{m+1}(r,k)}{B^m(r,k)} \times B^{m+1}(r,k)\right] \tag{6}$$

$$\text{TD\_entropy}(N,m,r,k) = -\ln\left[\frac{B^{m+1}(r,k)}{B^m(r,k)} \times B^{m+1}(r,k)\right] \tag{7}$$

根据 Pincus 建议,实验中将 m 设为 2,r 为 0.2SD,SD 为时间序列的标准差;同时将衰减系数 k 设为 0.01。

1.2 DFA 方法

给定一个时间序列 $X = \{x(i), i = 1,2,\cdots,N\}$,$N$ 是时间序列的长度,DFA 算法[16] 主要包含以下几个步骤:

(1) 运用以下公式,构造序列的轮廓序列。

$$Y(k) = \sum_{i=1}^{k} (x(i) - \bar{x}) \tag{8}$$

其中,$\bar{x}$ 表示序列 X 的均值,即 $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x(i)$;$Y(k)$ 表示序列的第 k 个值。

(2) 重构序列 $Y = \{Y(k), k = 1,2,\cdots,N\}$,将它划分成长度为 s 的互相不重叠的窗口,窗口个数为 $N_s = \text{int}(N/s)$。

(3) 分别在每个窗口 v 中,用最小二乘法拟合数据,则可获得局部趋势记为 $y_v(k)$。而拟合多项式的阶数可根据实际情况确定,主要用于去除线性的,二次的或者更高次的趋势。将去除趋势后的序列记作:

$$Y_s(k) = Y(k) - y_v(k) \tag{9}$$

(4) 计算去除趋势后每个窗口内的局部波动。

$$F_s^2(v) = \frac{1}{s}\sum_{k=1}^{s} Y_s^2[(v-1)s + k] \tag{10}$$

其中,$v = 1,2,\cdots,N_s$,表示正向顺序的窗口。

(5) 由于步骤 2 中 N 不一定能被 s 整除,故在序列尾部存在一小部分数据没有被分析。为了克服这个问题,将序列进行逆序操作,重复步骤 3、4,这样就得到 $2N_s$ 个窗口。

(6) 统计所有波动局部 $F_s^2(v)$ 来构造波动函数 $F(s)$。

$$F(s) = \left[\frac{1}{2N_s}\sum_{v=1}^{2N_s} F_s^2(v)\right]^{1/2} \quad (11)$$

对于给定的窗口长度值 s（或称为标度），可以得到对应的波动函数值 $F(s)$，设 $F(s)$ 与 s 具有幂律关系，即

$$F(s) \sim s^{\alpha} \quad (12)$$

其中，指数 α 称为标度指数。

根据式 12 可知 $\log F(s) \sim \alpha \log [s]$，则在 $(s, F(s))$ 的双对数图中，可以用最小二乘法拟合得到标度指数 α，可知 α 的数值大小可以反映序列的相关性，如表 1 所示。

表 1　标度指数 α 的含义

α 值	含义
α <0.5	序列具有负相关性或反记忆性，即过去增长(减小)趋势将意味着未来减小(增长)趋势，H 越接近 0，反记忆性就越强
α =0.5	序列中各个数据都是独立的，互不关联的，完全随机的，前一段时间的变化趋势不会对后面产生影响
0.5< α <1	序列具有正相关性或长记忆性，即过去一段时间的增长(减少)趋势将意味着未来相同时间间隔内有一个增长(减少)趋势，H 值越接近 1，长记忆性就越强
α =1	序列是 1/f 噪声
α >1	时间序列具有较强的非平稳性

2　数　据

实验数据来自国家海洋信息中心，卫星遥感中国近海的 SST 实测数据。数据观测点位于中国东部海域 130.125°E 经度上 31 个不同纬度。时间分辨率为 1 天，时间跨度为 1994 年 1 月 1 日至 2016 年 12 月 31 日，数据点长度为 8 401。主要选取了位于低纬度第 1 个观测点和中纬度第 16 个观测点以及高纬度第 31 个观测点的 SST 进行相关性及复杂性研究。

3　实　验

三个观测点的 SST 从 1994 年 1 月 1 日至 2016 年 12 月 31 日的时间序列如图 1 所示。可以看出，低纬度第 1 个观测点序列温差变化小于第 16 个观测点以及第 31 个观测点的温差变化，且第 1 个观测点的 SST 时间序列变化更复杂。第 16 个观测点以及第 31 个观测点的周期性更加明显。

接下来用二维熵度量这三个观测点每一年 SST 时间序列的复杂度，如图 2 所示。可以看出，低纬度第 1 个观测点每一年的 SST 时间序列的复杂度大于中纬度第 16 个观测点以及高纬度第 31 个观测点每一年的 SST 时间序列的复杂度。并且在 2007 年及以后三个观测点的 SST 时间序列的复杂度趋势都是降低的。

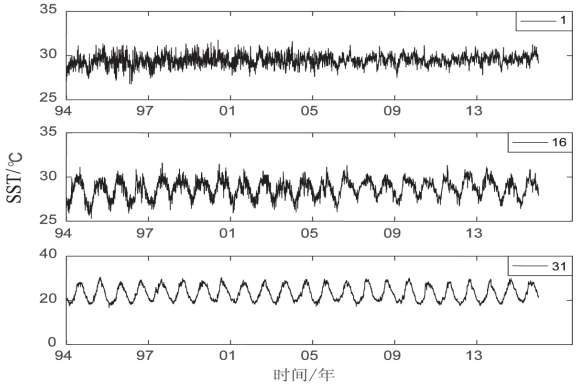

图 1　三个观测点的 SST 时间序列

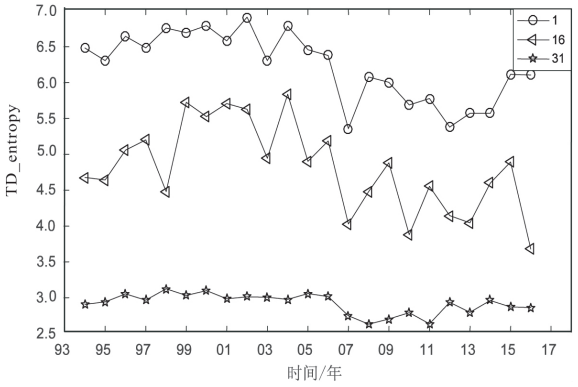

图 2　三个观测点每一年 SST 的复杂度

实际上从图 1 中可以看出第 1 个观测点 SST 时间序列的波动比较明显，相应的季节性等趋势不明显；而第 16 个观测点和第 31 个观测点的季节性都比较明显，相应的受其他因素影响的波动趋势就比较弱，序列看起来有规则，所以序列的复杂度相应就小。而低纬度区除了季节性会影响海表面温，其他因素，如太阳辐射和海洋大气热交换等影响更加明显。但是高纬度地区季节性因素相对来说对 SST 的影响性更大。

然后用 DFA 方法来研究三个观测点的长相关性。三个观测点每年的 SST 时间序列的 DFA 标度指数 α 如图 3 所示，每个观测点 23 年的 SST 时间序列的长相关性如图 4 所示。

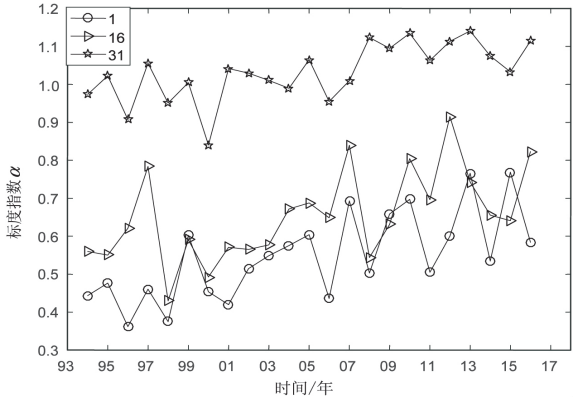

图 3　三个观测点每年 SST 时间序列的 DFA 标度指数

从图 3 可以看出，高纬度第 31 个观测点获得的 SST 时间序列每一年的 DFA 标度指数 α 比其他两个

观测点的标度指数都要高，且 α 值在 1 上下浮动，表明第 31 个观测点的 SST 时间序列具有较强的非平稳性和长记忆性。而第 1 个观测点的 SST 的 DFA 标度指数在 0.5 上下浮动，表明这个纬度每年的 SST 时间序列呈现出不同的特点，有时具有长记忆性，有时具有反记忆性。这个结果和复杂度结果相似，对于高纬度地区的 SST 的季节性比较明显，所以序列会呈现出长记忆性特征。而低纬度地区影响 SST 序列的因素比较多，当季节性因素影响较强时，序列会呈现出长记忆性；当其他因素影响更强时，序列会表现出随机性或者反记忆性。第 16 个观测点所在的中纬度地区的 SST 序列的季节性更加明显，同时其他因素对 SST 的影响也比第 31 个观测点强，所以第 16 个观测点 SST 的相关性在第 1 个观测点和第 31 个观测点之间波动。

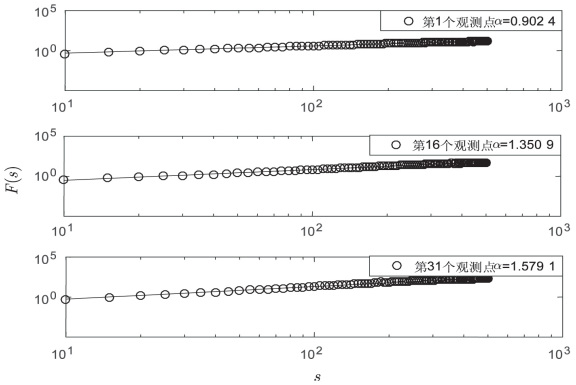

图 4 三个观测点 23 年的 SST 的 DFA 标度指数

图 4 是这三个观测点 23 年来(1994-2016 年)的 SST 时间序列的波动函数值 $F(s)$ 与窗口长度 s 之间的函数关系，直线是两者最小二乘法的拟合直线。可以看出，对于长时间序列，这三个观测点的 DFA 标度指数 α 都大于 0.5，且高纬度地区第 31 个观测点的 α 大于中纬度地区第 16 个观测点的 α 以及低纬度地区第 1 个观测点的 α。这三个观测点的 SST 长时间序列呈现出长记忆性且具有非平稳性。这是因为时间足够长时，序列的季节性特征就相对更加明显，序列的长期记忆性就强。

4 结束语

主要研究了海表面温(SST)时间序列的复杂性与相关性。根据一种改进的样本熵方法——二维熵方法来研究 SST 的复杂性；使用 DFA 方法研究 SST 的相关性。实验结果表明，周期性明显的高纬度地区的 SST 时间序列的复杂度低于低纬度地区 SST 时间序列的复杂性。因为影响低纬度地区的 SST 的因素较多，序列波动的周期性不明显，序列相对来说比较复杂。相对不复杂的高纬度地区 SST 时间序列长记忆性更加明显，且显示出较强的非平稳性。而相对复杂的低纬度地区 SST 时间序列呈现出反记忆性及长记忆性不同的特点。然而当序列的长度足够长时，不同纬度下的 SST 时间序列都呈现出明显的长记忆性及非平稳性。

参考文献：

[1] 陈艳秋，袁子鹏，王 元.SST 对黄海、渤海登陆热带气旋路径和强度的影响[J].海洋学报，2008，30(1)：31-41.
[2] RAO K G，GOSWAMI B N. Interannual variations of sea surface temperature over the Arabian Sea and the Indian Monsoon：a new perspective[J].Monthly Weather Review，2018，116(3)：558-568.
[3] 彭 婕.中国近海海表面温度日变化及其影响数值模拟研究[D].北京：国家海洋环境预报研究中心，2013.
[4] RICHMAN J S，MOORMAN J R.Physiological time-series analysis using approximate entropy and sample entropy[J].American Journal of Physiology Heart & Circulatory Physiology，2000，278(6)：H2039-H2049.
[5] 赵利民，朱晓军.基于局部均值分解与样本熵的脑电信号特征提取与分类[J].计算机工程，2017，34(2)：299-303.
[6] 成 娟，陈 勋，彭 虎.基于样本熵的肌电信号起始点检测研究[J].电子学报，2016，44(2)：479-484.
[7] COSTA M，GOLDBERGER A L，PENG C K.Multiscale entropy to distinguish physiologic and synthetic RR time series[J].Computers in Cardiology，2002，29：137-140.
[8] HURST H E.Long term storage capacity of reservoirs[J].Transactions of the American Society of Civil Engineers，1951，116：776-808.
[9] PENG C K，BULDYREV S V，GOLDBERGER A L，et al.Long-range correlations in nucleotide sequences[J].Nature，1992，356(6365)：168-170.
[10] PENG C K，BULDYREV S V，HAVLIN S，et al.Mosaic organization of DNA nucleotides[J].Physical Review E Statistical Physics Plasmas Fluids & Related Interdisciplinary Topics，1994，49(2)：1685-1689.
[11] 程 静，刘光远.基于情感心电信号的去趋势波动分析研究[J].西南大学学报：自然科学版，2016，38(2)：169-175.
[12] 李文菁.基于去趋势波动分析的中国气温变化趋势研究[D].湘潭：湘潭大学，2016.
[13] 李大夜.基于分形方法的金融市场长记忆性研究[D].北京：对外经济贸易大学，2017.
[14] 莫淑红，吕继强，沈 冰，等.基于去趋势波动分析的降雨演变特性研究[J].西安理工大学学报，2010，26(2)：148-151.
[15] 张美兰，金炜东，孙永奎，等.基于多重分形去趋势波动分析的高速列车运行状态识别方法[J].计算机应用研究，2015，32(10)：2978-2980.
[16] 夏佳楠.时间序列的多尺度不可逆性和复杂度研究[D].北京：北京交通大学，2017.