

互信息在海洋资料相关分析中的应用

郭衍游^{1,2}, 侯一筠¹, 刘战庆³

(1. 中国科学院海洋研究所, 山东 青岛 266071; 2. 中国科学院研究生院, 北京 100039; 3. 山东金洲矿业集团, 山东 威海 264501)

摘要: 为了揭示海洋资料的统计依赖特性, 提出了一种基于互信息的定性分析方法, 并对东海海表气温、气压及风速数据进行了分析。结果表明, 在数据资料较长时, 该方法能有效判定数据序列的线性、非线性相关特征, 特别适合于复杂的海洋数据资料分析。

关键词: 互信息; 相关分析; 非线性相关; 信息熵

中图分类号: P733

文献标识码: A

文章编号: 1000-3096(2004)10-0029-05

自从 1948 年 Shannon 提出“信息熵”概念以来, 信息论在诸多学科得到了广泛应用, 特别是由于它对非线性、复杂性现象有很强的分析能力, 现已成了研究海洋现象的一种重要方法。在海浪研究、台风预报及海气相互作用系统的非线性检测等方面引入“信息熵”均取得了良好的效果^[1~4]。互信息是信息论中的一个重要概念, 可用于分析两个随机变量之间的相互依赖程度, 相比自相关函数方法, 适用范围更广^[5~7]。由于海洋作用过程的复杂性和非线性, 通常是在一定条件下将其中的非线性因素简化为线性, 这就需要对这种假定进行检验, 如果非线性很强, 说明假定是不合理的, 不宜进行简化。作者提出了一种基于互信息的定性方法来揭示海洋资料中是否相互独立, 或是否存在线性相关或非线性相关, 并以东海海表气温、气压及风速资料为例进行了分析, 这对进一步研究其动力学性质是十分有益的。

1 基于互信息的相关分析方法

1.1 分析方法

统计学上, 两随机变量要么相互独立, 要么线性相关或者非线性相关。一个随机过程在不同时刻的系统状态就是两个随机变量, 下面提出的方法的目的就是定性地确定此两个变量之间的关系属于哪一种。

互信息(mutual information), 是包含于两不同随机变量之间平均共同信息量的度量^[8], 多个随机变量的互信息称作信息冗余(information redundancy)。设 X 和

Y 为两随机变量, 在 Y 已知的条件下, X 提供的新信息量可以用条件熵来表示^[9]。

$$H(X/Y) = H(X, Y) - H(Y) = - \sum_{ij} p(x_i, y_j) \log_a p(x_i, y_j) + \sum_j p(y_j) \log_a p(y_j) \quad (1)$$

其中 $H(X/Y)$ 是已知 Y 条件下 X 的条件熵, $H(X, Y)$ 是 X, Y 的联合熵, $p(Y_j)$ 是随机事件 Y_j 发生的概率, $p(X_i, Y_j)$ 是随机事件 X_i, Y_j 同时发生的概率, 常数 a 通常取 2, 10 或无理数 e 。记 $M(X; Y)$ 为 X 和 Y 的互信息, 则有:

$$M(X; Y) = H(X) - H(X/Y) = H(X) + H(Y) - H(X, Y) \quad (2)$$

易证, 当 X 与 Y 相互独立时, $M(X; Y) = 0$; 反之, $M(X; Y) > 0$ 。因此, 在统计学意义上, 互信息表达了两个随机变量之间的相互依赖性。

为便于不同随机变量互信息的比较, 对其进行标准化, 可以定义一个广义相关系数 ρ_g (generalized correlation coefficient)^[7]:

收稿日期: 2004-06-11; 修回日期: 2004-08-03

基金项目: 国家“863”资助项目(2001AA633070); 国家自然科学基金资助项目(40176010)

作者简介: 郭衍游(1977-), 男, 湖北阳新人, 博士, 主要从事海洋数据处理研究, 电话: 0532-2898516, E-mail: guoyanyou@126.com

$$\rho_g(X, Y) = \frac{M(X; Y)}{H(X, Y)} \quad (3)$$

显然 $0 < \rho_g(X, Y) < 1$, 且当 X 与 Y 完全相关时, $M(X; Y) = H(X) = H(Y) = H(X, Y)$, $\rho_g(X, Y) = 1$; 当 X 与 Y 完全独立时, $M(X; Y) = 0$, $\rho_g(X, Y) = 0$ 。

这样根据 $\rho_g(X, Y)$ 的值就可以判断 X 与 Y 是否相互独立。当 $\rho_g(X, Y)$ 值明显大于零时, 还需要进一步确定两者是线性相关还是非线性相关。Palus^[10,11]提出了一种基于互信息的辨别方法, 魏恩泊等^[5]将其应用于多维海洋资料的非线性检测。其原理是: 对于平稳且各态历经的随机过程 $X(t)$, 引入时间延迟 τ 构造两个变量 $X(t)$ 和 $X(t + \tau)$, 它们的互信息 $M(X(t); X(t + \tau))$ 是时间延迟 τ 的函数 $M(\tau)$, 可以由式(2)求得。现假设 $X(t)$, $X(t + \tau)$ 构成的二维随机变量符合二维正态分布, 即 $X(t)$ 为线性随机过程, 根据互信息的定义仍可计算出一个值, 记作 $L(\tau)$ 。如果 $X(t)$ 和 $X(t + \tau)$ 是线性相关, 那么 $M(\tau)$ 和 $L(\tau)$ 随时间延迟 τ 的变化规律必然非常相似的, 反之, 如果是非线性相关则明显不同。因此以时间延迟 τ 为 X 轴, 以 $M(\tau)$, $L(\tau)$ 为 Y 轴, 绘制 $M(\tau) - \tau$, $L(\tau) - \tau$

曲线, 比较两曲线的形状即可直观地判断序列是线性相关还是非线性相关。并且 $L(\tau)$ 可以方便地直接由二维正态变量的协方差矩阵 C 得到^[12]:

$$L(\tau) = \frac{1}{2}(\log_a C_{11} + \log_a C_{22}) - \frac{1}{2}(\log_a \sigma_1 + \log_a \sigma_2) \quad (4)$$

其中 σ_1, σ_2 为 C 的特征值, a 通常取 2, 10 或无理数 e 。

为验证上述方法的正确性, 对数值计算产生的不同类型的随机序列进行相关分析。

1.2 独立过程

随机过程理论表明, 均值为零的非周期性噪声 $X(t)$ 在时间延迟 τ 不太小时, $X(t)$, $X(t + \tau)$ 是相互独立的。用计算机产生的具有零均值的随机数序列 $x_i (i = 1, 2, \dots, n)$ 可以认为具有与该类噪声相同的性质, 理论上其广义相关系数 $\rho_g(\tau)$ 值应很小。计算结果显示, $\rho_g(\tau) < 0.05$, 接近零(图1), 表明该过程为独立过程, 与理论结果相吻合。此时, 由于已知该过程为独立, $M(\tau) - \tau$, $L(\tau) - \tau$ 曲线无实际意义。

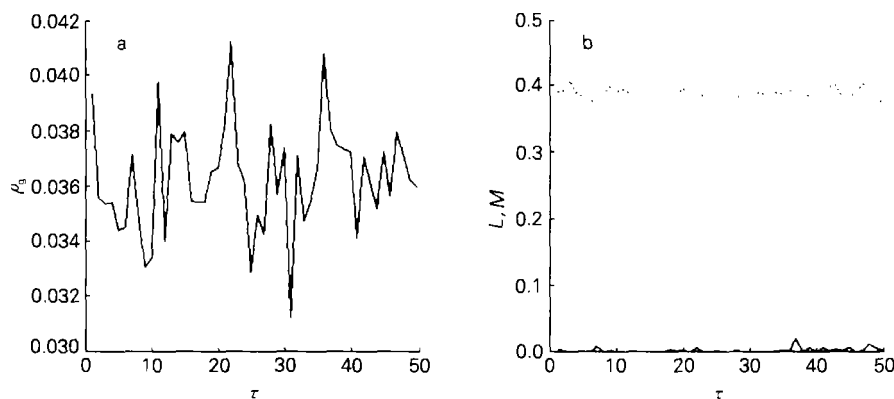


图1 独立过程的互信息特征

Fig. 1 Characters of mutual information in an independent process

a. $\rho_g(\tau)$; b. 实线为 $L(\tau)$, 虚线为 $M(\tau)$ (图2~图6同)

1.3 线性过程

用如下线性自回归方程可以产生一个线性时间序列:

$$x_t = \sum_{k=1}^{10} a_k x_{t-k} + \sigma e_t \quad (5)$$

其中 $a_k = 0, 0, 0, 0, 0, 0.19, 0.2, 0.2, 0.2, 0.2, k = 1, 2,$

$\dots, 10$, $\sigma = 0.01$, e_t 是均值为零, 方差为 1 的正态随机变量。

计算结果显示, $\rho_g(\tau)$ 大于 0.4 (图2a), 明显大于零, 也远大于上述独立过程的值, 表明该过程不同时刻的状态之间存在某种统计依赖性。因此, 进一步比较 $M(\tau) - \tau$, $L(\tau) - \tau$ 曲线形状, 可以看出两者变化趋势基本一致, 表明这种依赖特性是线性的 (图2b)。

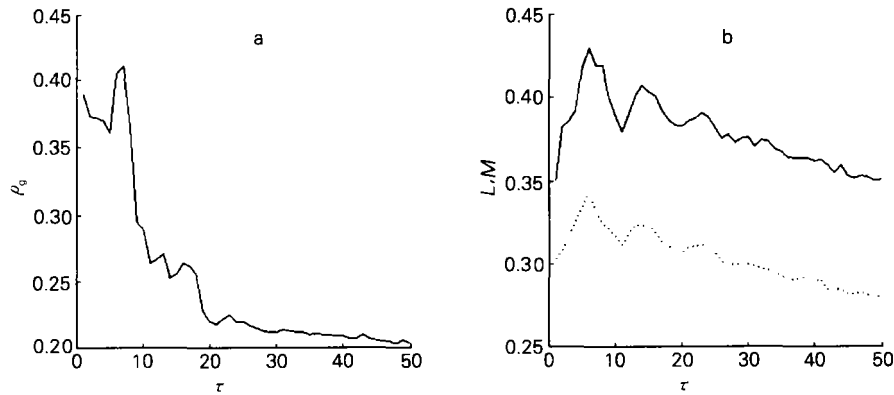


图2 线性过程的互信息特征
Fig. 2 Characters of mutual information in a linear process

1.4 非线性过程

根据著名的 Lorenz 方程:

$$\begin{aligned} & (dx/dt, dy/dt, dz/dt) = \\ & (10(y-x), 28x-y-xz, xy-8z/3) \end{aligned} \quad (6)$$

其中积分步长为 0.04, 初值为 (15.34, 13.68, 37.91), 可以得到一个典型的混沌时间序列。下面选取 X 变

量的时间序列数据进行计算。计算结果显示, $\rho(\tau) > 0.15$, 最大值接近 0.8 (图 3a), 表明该过程不同时刻的状态之间是统计依赖的。进一步比较曲线 $M(\tau) \sim \tau$, $L(\tau) \sim \tau$, 可以看出两者明显不相似, 表明该过程是非线性的 (图 3b)。

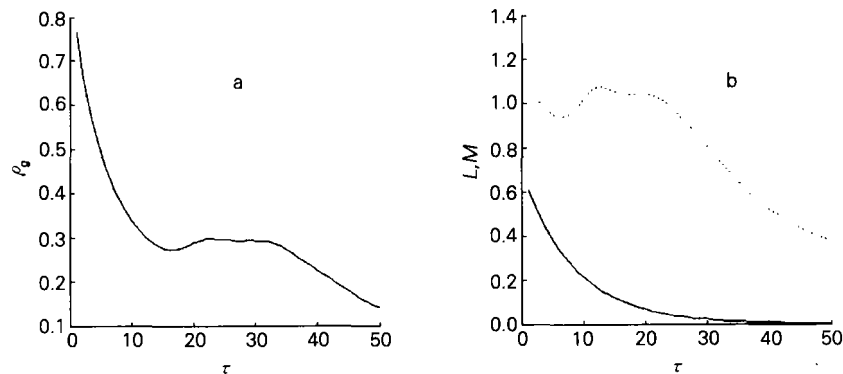


图3 非线性过程的互信息特征
Fig. 3 Characters of mutual information in nonlinear process

2 应用

在海洋资料处理中, 当数据长度较大时, 通常假设所获得的数据序列 $x(i=1, 2, \dots, n)$ 是平稳且各态历经的随机过程 $X(t)$ 的一次实现。在不同时刻 t_i, t_j , $X(t_i)$ 与 $X(t_j)$ 之间的相关特性只与时间延迟 $\tau = t_j - t_i$ 有关, 可以应用上述方法分析其相关特性。

作者选取的海平面气温、气压数据来自美国国家环境预报中心(NCEP) 1991~2001 年日平均再分析资料, 网格为 $2.5^\circ \times 2.5^\circ$, 中心位置地理坐标为 $(25^\circ \text{N}, 120^\circ \text{E})$; 海平面风速数据来源于东海浮标实测资料,

地理坐标为 $(36^\circ 40' \text{N}, 145^\circ 40' \text{E})$, 取样间隔 3 h, 时间范围为 1994 年全年。

分析结果为: 气温的广义相关系数 $\rho(\tau)$ 随时间延迟 τ 的增大而减小, 但始终维持在大于 0.15 这样一个较高的水平上 (图 4a), 表明气温的变化是不独立的。比较 $M(\tau) - \tau$, $L(\tau) - \tau$ 曲线, 时间延迟较小 ($\tau < 15$) 时两者变化趋势基本一致, 表明气温在时间延迟较小时存在较强的线性相关。在时间延迟增大 ($\tau > 20$) 时, 两曲线分离, $L(\tau)$ 继续线性衰减, 而 $M(\tau)$ 缓慢衰减并有微弱波动, 表现出明显的非线性相关特征 (图 4b)。

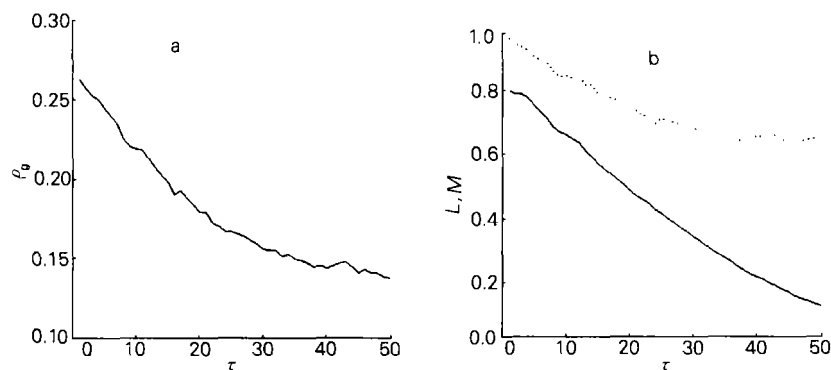


图 4 气温时间序列的互信息特征

Fig. 4 Characters of mutual information in air temperature time series

气压的广义相关系数 $\rho_g(\tau)$ 同样随时间延迟 τ 的增大减小,但在 $\tau < 30$ 时, $\rho_g(\tau)$ 一直维持在 0.1 以上 (图 5a), 表明气压的前后变化存在弱的相互依赖性。 $M(\tau) \sim \tau$, $L(\tau) \sim \tau$ 曲线形态非常相似说明这种依赖性是非线性的 (图 5b)。

风速的广义相关系数 $\rho_g(\tau)$ 随 τ 的增大呈不规

则变化,且 $\rho_g(\tau)$ 值非常小,在 0.03 与 0.045 之间波动 (图 6a), 表明不同时刻风速的相互依赖性非常弱,近于相互独立的。而 $M(\tau) \sim \tau$, $L(\tau) \sim \tau$ 曲线形态明显不同,前者随时间延迟增大呈不规则变化,后者接近零值基本不变 (图 6b), 说明这种微弱依赖性是非线性的。

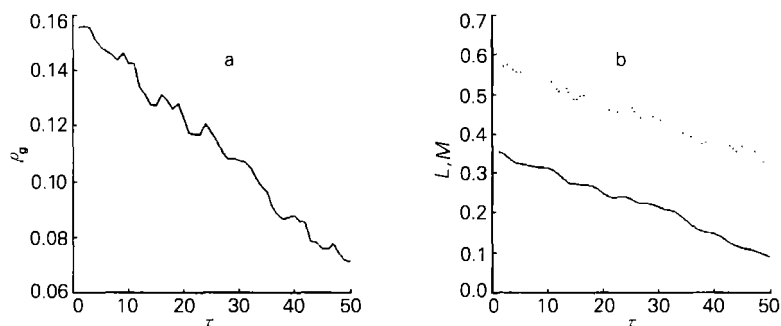


图 5 气压时间序列的互信息特征

Fig. 5 Characters of mutual information in air pressure time series

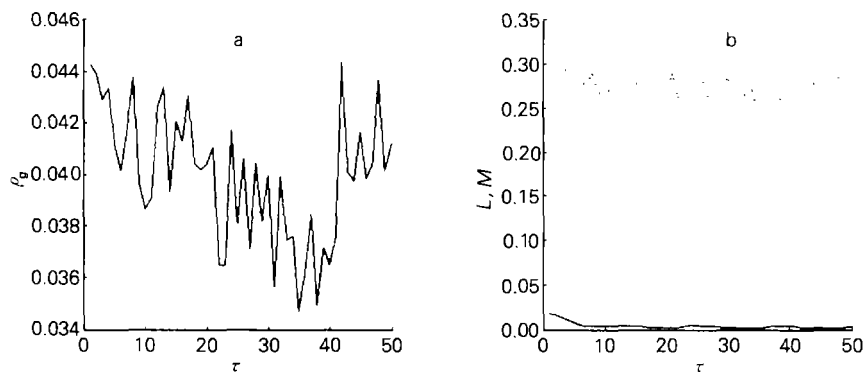


图 6 风速时间序列的互信息特征

Fig. 6 Characters of mutual information in a wind speed time series

3 小结

作者在互信息基础上分别对海表气温、气压和风速的相关性进行了分析。结果表明, 该方法既能处理线性相关, 又能处理非线性相关, 能很好适应海洋资料包含的非线性信息, 在海洋资料相关分析中取得了良好效果。

参考文献:

- [1] West B J. Deep Water Gravity Waves[M]. New York: Springer- Verlag, 1981. 249- 287.
- [2] 冯利华, 张萍名. 基于最大熵原理的台风统计预报[J]. 海洋科学, 2003, 27(3): 47- 51.
- [3] 吴克俭, 孙孚. 海浪波面的信息熵与海浪的谱宽度参量[J]. 海洋与湖沼, 1996, 27(3): 251- 257.
- [4] 吴克俭, 孙孚. 最大熵原理与海浪波高的统计分布[J]. 海洋学报, 1996, 18(3): 21- 26.
- [5] 魏恩泊, 宋金宝, 田纪伟. 南海多维海洋资料非线性检测[J]. 海洋与湖沼, 2003, 34(1): 109- 114.
- [6] Li Wentian. Mutual information functions versus correlation functions[J]. J Stat Phys, 1990, 60: 823- 837.
- [7] Tanaka N, Okamoto H, Naito M. Detecting and evaluating intrinsic nonlinearity present in the mutual dependence between two variables[J]. Phys D, 2000, 147: 1- 11.
- [8] 赵永龙, 丁晶, 邓育仁, 等. 以互信息为基础的广义相关系数及其应用[J]. 非线性动力学学报, 1997, 4(1): 9- 18.
- [9] 余成波. 信息论与编码[M]. 重庆: 重庆大学出版社, 2000. 12- 20.
- [10] Palus M. Information theoretic test for nonlinearity in time series[J]. Phys Lett A, 1993, 175: 203- 209.
- [11] Palus M. Testing for nonlinearity using redundancies: quantitative and qualitative aspects[J]. Physica D, 1995, 80: 186- 205.
- [12] Fraser A M. Reconstructing attractors from scalar time series: a comparison of singular system and redundancy criteria[J]. Physica D, 1989, 34: 391- 404.

Application of mutual information to ocean data correlation analysis

GUO Yan- you^{1,2}, HOU Yi- jun¹, LIU Zhan- qing³

(1. Institute of Oceanology, the Chinese Academy of Sciences, Qingdao 266071, China; 2. Graduate School of the Chinese Academy of Sciences, Beijing 100039, China; 3. Jinzhou Mineral Group, Weihai 264500, China)

Received: Jun , 11, 2004

Key words: mutual information; correlation analysis; nonlinear correlation; information entropy

Abstract: To reveal the statistic inter- dependence between different ocean data, this paper puts forth a qualitative method based on mutual information, and uses it to analysis the sea surface air temperature, air pressure and wind in East China Sea. Result shows that the method is effective to discriminate linear and nonlinear correlation characteristics in a comparatively long data series, and suitable for analysis of complex ocean data.

(本文编辑: 刘珊珊)