文章编号: 1000-0887(2008) 10-1227-10

应用数学和力学编委会, ISSN 1000-0887

奇异谱迭代插补的改进算法及其在缺损数据恢复中的应用

王辉赞[1,2,4], 张 韧[1,2], 刘 巍[3], 王桂华[4], 金宝刚[1,4]

(1. 解放军理工大学 气象学院 海洋与空间环境系, 南京 211101;
2. 中国科学院 大气物理研究所 大气科学和地球流体力学数值模拟国家重点实验(LASG), 北京 100029;
3. 西南交通大学 信息科学与技术学院, 成都 610031;
4. 国家海洋局 第二海洋研究所, 杭州 310012)

(郭兴明推荐)

摘要: 基于奇异谱分析(SSA)迭代插补的基本思想, 针对该方法在参数确定(主成分个数 K 和嵌入维数 M)上存在较大随意性和计算效率低等缺陷, 提出了一种参数优化的新算法 区间四分法, 该方法在误差曲线存在局部波动情况下仍能有效搜索到全局最优参数解, 且迭代插补精度和计算效率可得到显著提高 采用外逸长波辐射(OLR)逐日资料进行的插补试验对比分析表明, 基于区间四分法改进的 SSA 迭代插补方案的缺损数据恢复的效率和精度良好

关 键 词: 奇异谱分析; 外逸长波辐射; 缺损数据插补; 区间四分法
中图分类号: TP311;TP391 文献标识码: A

引 言

由于观测手段和观测环境的局限性, 在时空上均匀分布的连续历史或现实观测资料通常很难获得, 仅有的观测资料难以满足科学研究的需要, 因此如何对现有的缺损数据观测资料进行插补, 提取有效信息是挖掘和拓展数据信息资源的重要途径 国内外学者对缺损数据的插补进行了一些有意义的研究工作, 如回归函数方法、多项式插值、Kriging 插值、Kalman 滤波、最优插值、逐步订正、神经网络、分形插值和相空间重构预测等方法均被应用于缺损数据插补的算法研究之中, 取得了一定的效果

由于多元统计分析技术[1], 如经验正交函数(EOF)分析、主成分典型相关分析(PC-CCA)、奇异谱分析(SSA)、多通道奇异谱分析(MSSA)等, 能够揭示标量或向量场空间相关结构和时间演变规律, 因此在要素场的时间序列分析中得到了广泛应用 如何结合多元统计分析的思想, 对观测序列中的缺损数据点进行客观准确的插补和延伸具有较好的应用前景 江志红等

收稿日期: 2007-12-27; 修订日期: 2008-08-31
基金项目: 国家重点基础研究计划资助项目(2007CB816003); LASG 开放课题资助项目
作者简介: 王辉赞(1983), 男, 湖南浏阳人, 博士生(E-mail: wanghuizan@126.com);
张韧(1963), 男, 四川峨眉人, 教授, 博士, 博士生导师(联系人. Tel: +86-25-80831406; E-mail: zren63@126.com).

人[2-3]利用 PG-CCA 等方法对区域气温进行了插补试验, Beckers 和 Rixen[4] 和王桂华等人[5]分别利用 EOF 方法重构了 AVHRR 图像和太平洋海域三维温盐场, Kondrashov 和 Ghil[6-7] 利用 SSA 方法对海温、南方涛动指数等进行插补试验, 上述方法由于不用事先知道资料序列的先验估计误差信息, 不依赖于具体的计算模式, 因此具有较好的实用性 其中, SSA 和 MSSA 迭代插补方法对于事先未知物理本质的系统, 可以从它包括噪声的有限长度的观测序列中提取尽可能多的可靠信息, 并依据这些信息建立预报模型 因此在稀疏、散乱和缺损数据优化和恢复等方面表现出良好的应用前景 但是常规的 SSA 和 MSSA 迭代插补方法中主成分个数 K 和嵌入维数 M (也称窗宽) 等参数的选取具有较大的主观性和盲目性, 使得参数选取既不易搜索到全局最优解, 同时计算量也是巨大的 针对常规 SSA 和 MSSA 迭代插补方法存在的上述问题和不足, 本文提出了一种参数优化方法 区间四分法, 对常规 SSA 和 MSSA 迭代插补方法进行改进和完善, 并用该改进方法进行缺损数据插补实验和对比分析

1 SSA/MSSA 迭代插补的基本思想

SSA 主要针对单变量时间序列而言, MSSA 则是针对多变量时间序列而言, 它们都可以从时间序列中提取出其几个简单的主要模态表示其重要信息, 而滤去一些随机噪声 由于 SSA 和 MSSA 迭代插补方案思想类似, 故本文着重介绍单变量时间序列 SSA 迭代插补方案的基本思想

1.1 SSA 方法

设一维距平化的时间序列 $X(t) = \{x_t: t = 1, N\}$, 嵌入维数为 M SSA 方法可通过对角化时间序列 $X(t)$ 的落后协方差矩阵$(\boldsymbol{C}_X)_{M\ M}$ 来获得其谱信息 Broomhead 等人[8] 通过轨迹矩阵$(\boldsymbol{D})_{M\ N}$ 计算 $\boldsymbol{C}_X$, 其中 $N = N - M + 1$ $\boldsymbol{D}$ 是采用动力系统分数维估计的处理方法将序列时迟排列而成的相空间:

$$\boldsymbol{D} = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_{N-M+1} \\ x_2 & x_3 & x_4 & \cdots & x_{N-M+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_M & x_{M+1} & x_{M+2} & \cdots & x_N \end{pmatrix}, \tag{1}$$

则 $\boldsymbol{C}_X = \boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}/N$, 其中 $\boldsymbol{D}^{\mathrm{T}}$ 为 $\boldsymbol{D}$ 的转置矩阵 用 $D_{j,i}$ 表示轨迹矩阵 $\boldsymbol{D}$ 中第 j 行第 i 列的元素, 且有 $D_{j,i} = x_{i+j-1}$

由 $\boldsymbol{C}_X$ 得到其对应特征向量 $\boldsymbol{E}$ 和特征值 , 并将 $\boldsymbol{E}$ 及其对应的 按特征值的降序排列 $\boldsymbol{C}_X$ 的每个特征向量 $\boldsymbol{E}_k(1 \quad k \quad M)$ 有 M 个分量, 它反映了 $X(t)$ 序列的时间演变型, $\boldsymbol{E}_k$ 称为时间 EOFs(T-EOFs) $\boldsymbol{E}_k(1 \quad k \quad M)$ 和 $\boldsymbol{E}(j)(1 \quad j \quad M)$ 都具有正交性质 $\boldsymbol{C}_X$ 的特征值 $_k$ 解释了 $\boldsymbol{E}_k$ 方向的部分方差, 特征值之和给出了序列 $X(t)$ 的总方差 将时间序列作用于第 k 个 EOF 得到相关的主成分(T-PCs):

$$A_k(t) = \sum_{j=1}^{M} X_{t+j-1} E_k(j), \qquad 1 \quad t \quad N, \tag{2}$$

它表示 $\boldsymbol{E}_k$ 所表示的时间型在原序列的 $x_{i+1}, x_{i+2}, \cdots, x_{i+M}$ 时段占的权重

SSA 最重要的应用功能是通过重构成分实现的 根据特征向量 $\boldsymbol{E}$ 具有的正交性质, 可以利用 T-PCs 和 T-EOFs 的线性组合得到重构轨迹矩阵 $\boldsymbol{D}$:

$$D_{j,i} = \sum_{k} A_k(i) E_k(j), \qquad 1 \quad j \quad M; 1 \quad i \quad N, \tag{3}$$

其中, 为组成的特征成分的一个子集(即选取用来重构时间序列的前几个 $\boldsymbol{A}_k$ 和 $\boldsymbol{E}_k$ 的集合), $D_{j,i}$ 为重构轨迹矩阵 $\boldsymbol{D}$ 中第 j 行第 i 列的元素 于是, 根据时间序列元素 x_i 在轨迹矩阵 $\boldsymbol{D}$ 中的位置关系得知, 可通过式(5) 得到由 k 的主成分重构的时间序列 $X(t)$:

$$x_i = \begin{cases} \dfrac{1}{i}\sum_{j=1}^{i} D_{i,j}, & 1 \quad i \quad M-1, \\ \dfrac{1}{M}\sum_{j=1}^{M} D_{i,j}, & M \quad i \quad N, \\ \dfrac{1}{N-i+1}\sum_{j=i-N+M}^{M} D_{i,j}, & N+1 \quad i \quad N \end{cases} \tag{4}$$

$X(t)$ 是原时间序列 $X(t)$ 滤去了部分噪声并保留了较可靠周期信息得到的重构时间序列

MSSA 是针对多变量时间序列SSA 推广, MSSA 中通道数(即变量数) $L = 1$ 时的特例就是SSA 相对 SSA 而言, MSSA 的轨迹矩阵 $\boldsymbol{D}$ 是把各通道的单变量轨迹矩阵在列方向依次连接形成为 L M 行 N 列结构 由于利用 MSSA 的多变量时间序列重构方法与单变量时间序列重构方法类似, 故描述略去

1.2 SSA 迭代插补方案

SSA 迭代插补方案的核心思想: 1) 先进行内循环, 利用奇异谱分析提取时间序列的周期信息, 通过主成分重构使提取到的部分周期信息反映到待插补点上, 再对插补得到的新时间序列进行奇异谱分析, 以此类推不断迭代直至收敛(上述迭代方案的收敛性已从数学上得到证明[4]), 然后在外循环中增加主成分个数依次重复内循环过程; 2) 利用交叉验证方法优化嵌入维数 M 和需要选取的 SSA 主成分个数 K

SSA 迭代插补方案算法流程主要分如下两步:

第 1 步 用交叉验证方法确定最优参数 M 和 K

a) 给参数赋初值 $M = 1, K = 1$, 并给出 $M_{\max}$;

b) 将原时间序列 $X(t)$ 中的数据分 3 类: 训练数据 X_{train}、交叉验证数据 $X_{\text{cross-valid}}$ 和待插补数据 X_{fill} 其中, X_{train}、$X_{\text{cross-valid}}$ 为已知数据(或称观测数据), 但 $X_{\text{cross-valid}}$ 是从已知数据中随机选择的, 在插补过程中被看作未知数据进行插补, 以便最后利用 $X_{\text{cross-valid}}$ 处的插补值与该处已知数据值进行比较检验插补效果, X_{fill} 为待对其进行插补的缺测数据;

c) 令 $n = 0$, 将 X_{train} 去中心化并记录其平均值 X_{ave}、$X_{\text{cross-valid}}$ 和 X_{fill} 处用 0 填补, 得到 $X_n(t)$;

d) 对时间序列 $X_n(t)$ 按 1.1 节的介绍进行嵌入维数为 M 的奇异谱分析 取前 K 个主要成分得到重构时间序列 X_{recon}, $X_n(t)$ 中 $X_{\text{cross-valid}}$、X_{fill} 处的值用 X_{recon} 中对应位置值替代, 得到 $X_{n+1}(t)$;

e) 如果 $\max|X_{n+1}(t) - X_n(t)|$, 则返回 f); 否则, 令 $n = n+1$, 返回 d);

f) $X_{n+1}(t) = X_{n+1}(t) + X_{\text{ave}}$, 计算 $X_{n+1}(t)$ 在 $X_{\text{cross-valid}}$ 处的插补值与该处已知观测数据值的均方根误差 $e_r(M, K)$;

g) 如果 $K < M$, 则 $K = K+1$, 跳至 ;

如果 $K = M$, 则 $M = M+1, K = 1$;

如果 $M = M_{\max}$, 则返回 h); 否则, 返回 b) 开始新的插补过程;

h) 找出使得均方根误差达到最小的 M 和 K, 取为最优参数 M_{opt} 和 K_{opt}, 程序结束

第 2 步　SSA 插补缺测数据

a) 将原时间序列分为两类: 训练数据 X_{train} 和待插补数据 X_{fill};

b) 令 $n = 0$, 将 X_{train} 去中心化并记录其平均值 X_{ave}, X_{fill} 处用 0 填补, 得到 $X_n(t)$;

c) 对时间序列 $X_n(t)$ 进行嵌入维数为 M_{opt} 的奇异谱分析, 取前 K_{opt} 个主要成分得到重构时间序列 X_{recon}, $X_n(t)$ 中 X_{fill} 处的值用 X_{recon} 中对应位置值替代, 得到 $X_{n+1}(t)$;

d) 如果 $\max|X_{n+1}(t) - X_n(t)|$, 则 $X_{n+1}(t)$ 为插补后的序列, 程序结束; 否则, 令 $n = n + 1$, 返回 c)

需要注意, 由于交叉验证数据位置的选取是从已知数据中随机产生的, 为了使得交叉验证结果更加稳定, 在第 1 步中可以对给定后的 M 和 K 进行多组实验求取多组均方根误差的平均值

MSSA 迭代插补算法流程与 SSA 基本相同, 主要区别在于MSSA 方案第 1 步 d) 和第 2 步 c) 进行的是多通道奇异谱分析, 第 2 步 g) 中 K 满足的条件不再是 K　M, 而是 K　$\min(M$　L, $N)$, 其具体算法流程在此不做赘述

1.3　常规的参数选取方法及不足

由 1.2 节可知, SSA 迭代插补方案主要分两步, 即用交叉验证法确定最优参数和确定最优参数后的 SSA 插补缺测数据　但该方案存在的问题是, 由于第 1 步采取的是类似于穷举法逐个循环 M 和 K 来确定最优参数, 对运算过程耗时较多, 而对插补数据来说, 第 1 步的意义仅在于为第 2 步提供最优参数, 因此, 需针对最优参数确定方法存在的不足进行改进

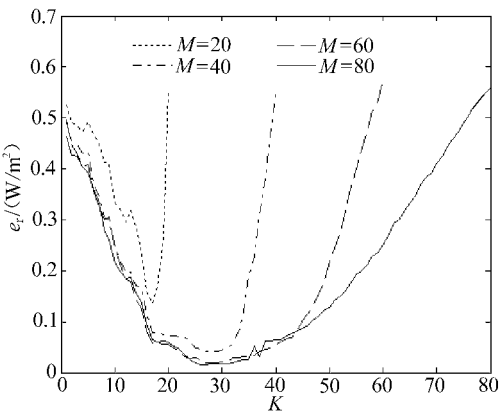

图 1　取不同嵌入维数 M 时误差随 K 变化曲线图

首先, 分析均方根误差函数 $e_r(M, K)$ 对 M 和 K 的依赖性　主成分个数 K 选取的大小反映的是序列周期信号的多少　当嵌入维数 M 不变时, 如果选取的主成分个数 K 较小, 则不足以反映序列的主要周期性信息; 但如果选取的主成分个数 K 较大, 则由于将参杂了较多的噪声信息也反映到了新的时间序列中, 使得在插补点处的主要周期信息被噪声淹没, 从而造成重构得到的时间序列变化很小　特别地, 当 K 取值为 M 时, 主要周期和噪声都得以保留, 重构的时间序列与原时间序列完全相同, 也就失去了重构的意义　所以, 当 M 不变时, K 选取较大或较小都会使得误差函数值较大, 应该在中间位置存在某个 K 值使得误差函数达到最小　经过多组数据插补实验也证实了这一点结论, 在下面的插补实验中也可以看出这一点　同时通过分析和实验还可以发现, 误差函数中 M 一定时, $e_r(M, K)$ 随 K 变化的误差曲线总体变化趋势是先减小, 至最小值后再增大, 但对于细节而言并非严格, 而是存有小的波动, 振幅不大　图 1 为 SSA 迭代插补实验一个个例的误差曲线图, 十分具有代表性, 从图上很容易看出误差函数 $e_r(M, K)$ 对 K 的依赖关系　同时, 从图中也可以发现, $e_r(M, K)$ 对 M 的依赖关系不如 K 那么明显的特征, 也并非 M 越大时对应的误差最小值越小

其次, 改进 SSA 迭代插补方案中对于最优参数的选取方法　在 1.2 节算法中, 最优参数的选取是采用对 M 和 K 逐个循环的方法, 使得运算量较大　因为误差函数对 M 的依赖关系特

征不明显, 所以主要针对 M 一定时, K 的最优选取 方法进行改进 但由于误差曲线存在局部波动的特性, 使得采用 固定 M 不变, K 逐次增大, 依次比较 $K = i+1$ 与 $K = i$ 误差值大小, 如果 $e_r(M, i) < e_r(M, i+1)$ 时的 i 作为该 M 值时的最优 K 的方法会陷入第一个局部最优值而出现错误

基于上述分析, 常规 SSA 和 MSSA 的参数确定方法存在着较大的随意性和盲目性, 进而直接影响缺损数据的插补质量和计算效率 为此, 结合误差曲线的总体变化趋势和局部存在波动的特性, 本文提出一种改进的最优参数选取方法 求误差函数最小值的区间四分算法

2 SSA/MSSA 迭代插补的改进算法 区间四分法

针对上述方法存在的不足, 我们提出 SSA/MSSA 迭代插补的改进算法 区间四分法 由于本算法对于类似问题具有普适性, 在此不妨设函数关系式为 $y = f(n)$, 其自变量为整数为表现函数单调增加和减小的性质, 不失一般性, 设函数满足 总体变化趋势先减小, 后增大, 中间存有小波动 的性质, 旨在求取函数在整数区间 $[a_1, e_1]$ 上的最小值; 该算法对函数单调增加或减少的情况也同样适用

2.1 区间四分法的基本思想

将区间 $[a_1, e_1]$ 平均分成 4 个小区间, 区间划分点分别为 b_1、c_1、d_1, 再判断最小值属于哪一个小区间或哪两个相邻的小区间; 再把缩小范围后最小值所在的小区间再细分为 4 个小区间, 再判断最小值所在的更小区间, 等等; 重复这一过程, 直至区间变得不能再细分 这时, 计算小区间内每个点的值, 并从中找出最小值点所在位置

那么如何判断最小值属于哪一个小区间或哪两个相邻的小区间? 设 4 个子区间的 5 个端点值依次为 a_i、b_i、c_i、d_i、e_i, 且对应函数值分别为 f_a、f_b、f_c、f_d、f_e, 那么可得: 1) 当 $f_b \quad \min(f_a, f_c)$ 时, 最小值一定位于 $[a_i, c_i]$ 区间; 2) 当 $f_c \quad \min(f_b, f_d)$ 时, 最小值一定位于 $[b_i, d_i]$ 区间; 3) 当 $f_d \quad \min(f_c, f_e)$ 时, 最小值一定位于 $[c_i, e_i]$ 区间 如果这 3 种条件都不满足, 则一定有 $f_b > \min(f_a, f_c), f_c > \min(f_b, f_d), f_d > \min(f_c, f_e)$ 同时成立

若 $f_a < f_b$ 成立, 则根据函数 $y = f(n)$ 性质(具有 先减小, 后增大 或者 单调 特征), 一定有 $f_a < f_b < f_c < f_d < f_e$, 最小值一定落在 $[a_i, b_i]$; 若 $f_a > f_b$ 成立, 则根据 3 个条件不等式 $f_b > \min(f_a, f_c)$、$f_c > \min(f_b, f_d)$、$f_d > \min(f_c, f_e)$ 一定可以推出 $f_a > f_b > f_c > f_d > f_e$, 最小值一定落在 $[d_i, e_i]$ 因此只要在原来 3 个判断条件基础上, 增加两个条件: 4) 当 $f_b < f_c$ 时, 最小值一定落在 $[a_i, b_i]$ 和 5) 当 $f_c > f_d$ 时, 最小值一定落在 $[d_i, e_i]$, 就把各类情况都考虑在内

区间四分法算法的优点主要有两点: 1) 区间四分法算法相比原方法寻找最优参数所用时间明显减小 原算法的时间复杂度为 $O(n)$, 而区间四分法的时间复杂度为 $O(\log_2 n)$; 2) 不容易陷入局部极小值 因为采用的是区间逐步缩小搜索法, 而函数只存在小波动, 所以对于较大区间来说, 小波动不影响区间的正确寻找 当区间缩小到一定程度(即不能再细分)时, 采用计算小区间内所有点函数值并从中找最小值策略, 可避免了小波动的影响

2.2 区间四分法的算法步骤

a) 令 $i = 1$, 给 a_1、e_1 赋值(表示在 $[a_1, e_1]$ 范围内寻找函数最小值); 计算 $f_a = f(a_1), f_e = f(e_1)$, flag$C = 0$(标记中间点 C 是否需要重新计算, 是则标记为 0);

b) 令 $= [(e_i - a_i)/4]$; 如果 1, 则

$b_i = a_i +$, $d_i = e_i -$, $f_b = f(b_i), f_d = f(d_i)$;

如果 flag$C$ = 0, 则 c_i = a_i + 2 , f_c = $f(c_i)$; 如果 < 1, 则计算 a_i + 1, a_i + 2, , e_i − 1 处的函数值, 从 a_i, a_i + 1, , e_i − 1, e_i 中找出其最小值, 程序结束;

c) 如果 f_b min(f_a, f_c), 则 a_{i+1} = a_i, c_{i+1} = b_i, e_{i+1} = c_i, f_e = f_c, f_c = f_b, flag$C$ = 1, 返回 d);

如果 f_c min(f_b, f_d), 则 a_{i+1} = b_i, c_{i+1} = c_i, e_{i+1} = d_i, f_a = f_b, f_e = f_d, flag$C$ = 1, 返回 d);

如果 f_d min(f_c, f_e), 则 a_{i+1} = c_i, c_{i+1} = d_i, e_{i+1} = e_i, f_a = f_c, f_c = f_d, flag$C$ = 1, 返回 d);

如果 f_b < f_c, 则 a_{i+1} = a_i, e_{i+1} = b_i, f_e = f_b, flag$C$ = 0, 返回 d);

如果 f_c > f_d, 则 a_{i+1} = d_i, e_{i+1} = e_i, f_a = f_d, flag$C$ = 0, 返回 d);

d) i = i + 1, 返回 b)

下面以图 2 误差曲线为例, 对区间四分法进行说明

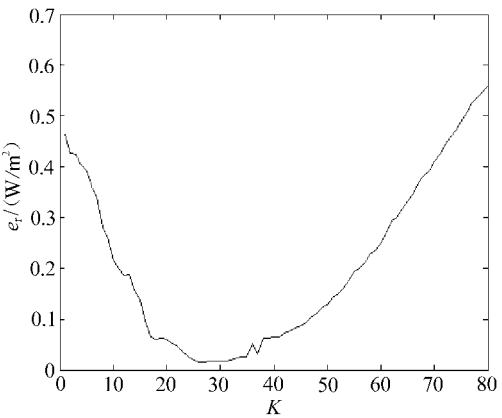

图 2　误差曲线图

可以看出, 图 2 的误差曲线随 K 的总体变化趋势是先减小后增大且存在小波动　运用区间四分法算法对其误差最小值进行搜索, 区间缩小过程表现为:

a_1 = 1, b_1 = 20, c_1 = 39, d_1 = 61, e_1 = 80, 对应值: 0.464 3, 0.060 5, 0.061 9, 0.266 9, 0 560 2;

a_2 = 1, b_2 = 10, c_2 = 20, d_2 = 30, e_2 = 39, 对应值: 0.464 3, 0.215 0, 0.060 5, 0.018 0, 0 061 9;

a_3 = 20, b_3 = 24, c_3 = 30, d_3 = 35, e_3 = 39, 对应值: 0.060 5, 0.025 8, 0.018 0, 0.026 5, 0 061 9;

a_4 = 24, b_4 = 26, c_4 = 30, d_4 = 33, e_4 = 35, 对应值: 0.025 8, 0.014 9, 0.018 0, 0.022 6, 0 026 5;

a_5 = 24, b_5 = 25, c_5 = 26, d_5 = 29, e_5 = 30, 对应值: 0.025 8, 0.019 1, 0.014 9, 0.017 1, 0.018 0;

a_6 = 25, b_6 = 26, c_6 = 26, d_6 = 28, e_6 = 29, 对应值: 0.019 1, 0.014 9, 0.014 9, 0.016 5, 0.017 1;

a_7 = 25, e_7 = 26, 对应值: 0.019 1, 0.014 9

于是, 找到最小值位置: K = 26; 对应最小值: 0.014 9

实验证明, 对于存在小波动的误差曲线(该误差曲线存在 5 个极小值点, 分别位于 K = 12, 18, 26, 31, 37), 利用区间四分法运算速度比原循环算法快很多(原算法需计算 80 个点的误差函数值, 而区间四分法只需计算 15 个点的误差函数值), 且不易受小波动影响, 能找到全局最小值点 K = 26

区间四分法对于 MSSA 迭代插补方案改进方法和效果与 SSA 相同, 故相应的描述省略

3 SSA/MSSA 迭代插补改进算法的应用试验

3.1 资料介绍

选取由 NCEP/NCAR 提供的逐日 OLR(外逸长波辐射)数据资料,单位为瓦每平方米(W/m^2),其网格精度为 2.5 2.5,资料区域范围为:(90 E~ 140 E,10 S~ 30 N),时间序列长度为:2004.5.1~ 2006.4.30(共 730 d),共有 730 357= 260 610 个网格点数据

3.2 OLR 资料插补试验

3.2.1 迭代方案说明

SSA 迭代插补方案是针对单变量时间序列而言,MSSA(多通道奇异谱)迭代插补方案可以处理多变量时间序列问题 因此,将空间网格点拉伸,对 OLR 数据进行插补实验,在参数选取时采用本文提出的区间四分法 从 260 610 个网格点数据中随机抽取 40% 的数据作为预测数据点,余下的 60% 作为已知数据 用已知数据其中 10%(占总数据的 6%)作为交叉验证数据,而剩下的 90%(占总数据的 54%)作为训练数据进行 SSA 迭代插补试验 任意选取某次 SSA 迭代插补过程中迭代解的收敛曲线图表明,迭代方案的解严格收敛且速度较快,迭代有限次数完全能达到收敛条件(图略)

对参数 M 依次取值 1, 2, ,然后按照本文提出的改进参数选取法(区间四分法)对其进行最优参数 K 的搜索,可以根据交叉验证数据的均方根误差,依次确定不同 M 值对应的最优参数:$M=1, K=32$;$M=2, K=52$;$M=3, K=68$; 同时得到不同 M 值时对应的均方根误差(见表 1)

表 1 M 为不同值时选取的最优参数 K 交叉验证效果对比

最优参数值	相关系数 R	交叉验证均方根误差 e_r/(W/m^2)
$M=1, K=32$	0.837 80	22.864
$M=2, K=52$	0.851 20	21.746
$M=3, K=68$	0.850 76	21.880

注 $M=1$ 时即为 EOF 迭代插补方案最优参数;$M=2$ 时即为 MSSA 迭代插补方案最优参数

从表 1 中分析可以看出,当参数 $M=2$ 时对应的最优参数 $K=52$ 的插值效果最好,因为 $M=3$ 插值效果开始变差,所以不用继续对 $M=3$ 之后的值进行最优参数的搜索 需要指出的是,EOF 是 MSSA 中嵌入维数 $M=1$ 的特例,所以 MSSA 迭代插补方案取 $M=1$ 时对应的最优参数 $K=32$ 就是 EOF 迭代插补方案的最优参数 表中数据说明,MSSA 迭代插补方案选择的参数范围更广,能够得到比 EOF 更优的参数,MSSA 迭代插补方案比 EOF 迭代插补方案更具优越性

3.2.2 插补试验效果分析

3.2.2.1 MSSA 迭代插补方案效果分析及其与 EOF 方案对比

图 3 给出了 MSSA 迭代插补方案在所有缺损

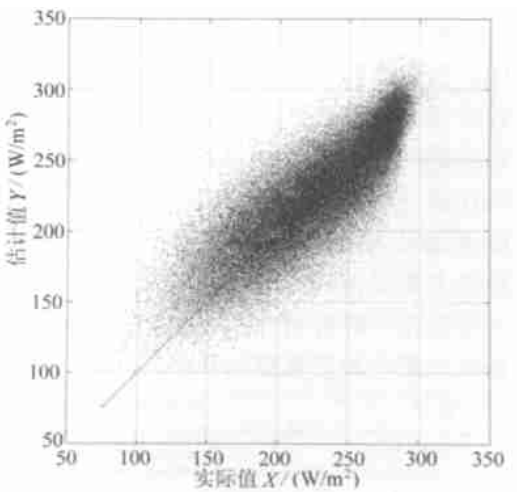

图 3 缺损点的 OLR 实际值和估计值比较

点(共 104 244 个点) 处, OLR 的实际值和用 MSSA 迭代插补方案得到的估计值比较图(用横坐标表示实际值 X, 纵坐标表示估计值 Y), 图中对应点基本都集中在直线 $Y = X$ 附近, 表明该插补方案能得到很好的插补效果, 插补值与实际值的相关系数可达到 85. 53%, 均方根误差为 21 483 W/m^2

a) 为检验 OLR 空间场的插值效果, 从时间序列场共 730 d(2006. 4. 30~ 2004. 5. 1) 中任意抽取 2 d, 得到实际场与插补方法恢复得到的 OLR 场的对比 结果表明, EOF 迭代插补方法不失为当前比较好的一种插补方法, 对于缺失达 40% 数据的空间场仍能较好的恢复, 能达到了较高的精度 它与 MSSA 迭代插补方案插补分析结果的形势场十分相似, 但在细节上插补仍有欠缺 EOF 迭代方案的插补效果略逊于 MSSA 迭代方案(见表 2)

b) 为检验 OLR 单点时间序列插值效果, 不妨以[90 E, 30 N] 点格点的时间序列(共 730 个时间点) 为例进行插值效果比较

插补方案的对比试验结果表明: MSSA 迭代插补方法较 EOF 迭代插补方法效果更佳, 后者的插补结果与真实数据序列相关系数为 0. 775 0; 而前者插补曲线与真实数据序列相关系数达 0. 825 1

为了进一步定量将 EOF 迭代插补方案与 MSSA 迭代插补方案的效果进行对比, 计算序列的所有缺损点插补值与实际值的相关系数及其均方根误差(表 2) 表中可以看出, $M = 2$(即 MSSA 迭代插补方案) 比 $M = 1$(即 EOF 迭代插补方案) 的均方根误差要小, 相关系数要高(因缺失点达 104 244 个, 故具有很好的统计意义), 因此, MSSA 的插补效果较 EOF 更好

表 2 M 为不同值时选取的最优参数 K 插补效果对比

不同 M 值的最优参数 K	相关系数 R	缺损点估计均方根误差 e_f/(W/m^2)
$M = 1, K = 32$	0. 845 78	22. 268
$M = 2, K = 52$	0. 855 30	21. 483
$M = 3, K = 68$	0. 853 75	21. 583

注 $M = 1$ 时即为 EOF 迭代插补方案最优参数; $M = 2$ 时即为 MSSA 迭代插补方案最优参数

3. 2. 2. 2 MSSA 迭代插补改进算法与 MSSA 常规算法的对比

MSSA 迭代插补改进算法较常规 MSSA 方案的优势主要体现在计算时间(见表 3, 给出了 OLR 迭代插补中两者的比较) 和计算精度上 从表中可以清楚地看出, 改进后的迭代插补方案较常规方案的计算时间有极为显著的提高, 计算速度较常规方法提高数 10 倍, 特别是对于大数据量的计算显示出比常规方案更明显的优势 因为常规 MSSA 迭代插补方案需要较大的时间代价, 所以在利用其进行实际插补操作时常常通过取大 K 的时间步长来减少运算时间, 但由于这无法搜到全局最优参数 K, 从而使得插补方案的精度较低 而由于改进的 MSSA 迭代插补方案能搜到 K 的全局最优值, 使得插补后的数据具有较高的精度和准确性 因此, 区间四分法是一种针对 SSA/MSSA 迭代方案非常有效的方法, 有助于改进和发挥 SSA/MSSA 迭代插补方案在缺损数据插补中的作用和优势

表 3 常规的与改进的 MSSA 迭代插补方案计算时间比较

参数 M 取值	常规 MSSA 迭代需计算 K 个数	改进的 MSSA 迭代需计算 K 个数	速度提高倍数
1	357	19	17. 8
2	714	22	31. 5
3	728	22	32. 1

4 小 结

SSA 和 MSSA 迭代插补方法是一种新颖且有很大应用前景的缺损数据插补方法. 无论对单变量时间序列, 还是多变量时间序列资料的空间信息缺损和时间信息缺损都能够尽量充分地反映时间序列的周期信息和滤掉噪声, 对缺损数据取得较好的恢复效果. MSSA 迭代插补方案实际是 EOF 迭代插补方案的扩展. 当 MSSA 中嵌入维数 $M = 1$ 时, 就是 EOF 方案, MSSA 迭代插补方案要优于 EOF 迭代插补方案.

常规 SSA/MSSA 迭代插补的参数选取方法存在一定的盲目性和人为性, 且计算效率较低. 针对该问题, 本文提出了其改进算法——参数优化区间四分方法. 该方法能够有效改进和提高常规 SSA/MSSA 迭代插值方法计算效率和计算精度. 其优势主要表现为: 1) 在误差曲线存在局部波动的情况下仍能有效搜索到最优参数解; 2) 显著提高了迭代插补的计算速度和效率(常规算法时间复杂度为 $O(n)$, 本文提出的区间四分法时间复杂度为 $O(\log_2 n)$).

研究发现, MSSA 算法还存在如下方面不足: MSSA 虽然能够对空间数据点场组成的时间序列进行插补(将空间信息看成是多个变量), 能够将时空两个方向的信息都反映出来, 但不能对无任何数据记录的点进行插值加密. 换句话说, 如果某一空间点地理位置已知, 但无任何时间序列资料数据, 那么 MSSA 迭代插补算法对于该点数据的插补将无能为力, 因为该算法中没有真正考虑二维空间信息, 而只是考虑了空间上的相互次序信息. 如何改进这个不足, 我们认为可以从以下 3 方面入手: 一是可以引入空间点地理位置信息矩阵, 使其能将空间点距离等信息也能反馈到 MSSA 奇异谱分析中去; 二是通过对 MSSA 进行维数扩展分解, 如扩展到三维等, 这样更多的空间信息就可反映出来; 三是将 MSSA 迭代插补方案与其它能反映空间地理信息的插值方法结合起来, 如 Kriging 插值等. 上述改进思想和算法实现是我们下一步工作中拟研究和探索的目标.

[参 考 文 献]

[1] 吴洪宝, 吴蕾. 气候变率诊断和预测方法[M]. 北京: 气象出版社, 2005.

[2] 江志红, 丁裕国, 屠其璞. 基于 PC-CCA 方法的气象场资料插补试验[J]. 南京气象学院学报, 1999, **22**(2): 141-148.

[3] 江志红, 丁裕国, 屠其璞. 气象场序列几种插补方案的对比试验[J]. 南京气象学院学报, 1999, **22**(3): 352-359.

[4] Beckers J M, Rixen M. EOF calculations and data filling from incomplete oceanographic datasets [J]. Journal of Atmospheric and Oceanic Technology, 2003, **20**(12): 1839-1856.

[5] 王桂华, 刘增宏, 许建平. 利用 Argo 资料重构太平洋三维温盐场和流场[A]. 见: 许建平 主编. Argo 应用研究论文集[C]. 北京: 海洋出版社, 2006, 16-25.

[6] Kondrashov D, Ghil M. Spatio-temporal filling of missing points in geophysical data sets[J]. Nonlinear Processes in Geophysics, 2006, **13**(2): 151-159.

[7] Kondrashov D, Ghil M. Reply to T Schneider s comment on "Spatio-temporal filling of missing points in geophysical data sets" [J]. Nonlinear Processes in Geophysics, 2007, **14**(1): 3-4.

[8] Broomhead D S, King G P. Extracting qualitative dynamics from experimental data[J]. Physica D, 1986, **20**(2/3): 217-236.

Improved Interpolation Method Based on Singular Spectrum Analysis Iteration and Its Application in Missing Data Recovery

WANG Hui-zan[1,2,4], ZHANG Ren[1,2], LIU Wei[3],
WANG Gui-hua[4], JIN Bao-gang[1,4]

(1. Institute of Meteorology, PLA University of Science and Technology, Nanjing 211101, P.R. China;
2. The State Key Laboratory of Numerical Modeling Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, P.R. China;
3. School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, P.R. China;
4. Second Institute of Oceanography, State Oceanic Administration, Hangzhou 310012, P.R. China)

Abstract: A novel algorithm called interval quartering algorithm was proposed to improve the insufficiency of the conventional singular spectrum analysis iterative interpolation on parameter selection (including the number K of principal component and the embedding dimension M). Based on the improved singular spectrum analysis iterative interpolation, the interpolated test and comparative analysis was carried out to the outgoing longwave radiation daily data. The results show that interval quartering algorithm can not only find the global optimal parameter to the error curve which has local oscillation effectively but also has the advantage of fast computing speed, and this improved interpolation method is very effective to the interpolation of missing data.

Key words: singular spectrum analysis; outgoing longwave radiation; interpolation of missing data; interval quartering