第 28卷 第 6期
2009年 12月

海 洋 环 境 科 学
MARINE ENVIRONMENTAL SCIENCE

Vol. 28, No. 6
Dec. 2009

高维空间插值在海洋环境数据预处理中的应用*

李 峋[1], 仵彦卿[1], 范海梅[2]
(1. 上海交通大学 环境科学与工程学院, 上海 200240; 2. 国家海洋局东海环境监测中心, 上海 200120)

摘 要: 基于国家海洋局东海分局环境监测中心 1999到 2006年在上海海域的原始监测数据, 提出了一种新的数据预处理方法。该方法在高维空间中使用最邻近插值算法, 消除了周期性变化以及边界截断误差的影响, 实现了对研究区域任意时间, 任意地点的插值, 在海洋环境领域有着一定的应用前景。

关键词: 最邻近插值; 高维空间; 数据预处理; 海洋环境
中图分类号: Q141 文献标识码: A 文章编号: 1007-6336(2009)06-0729-05

Application of high-dimension interpolation method in pretreatment of marine environmental data

LI Xun[1], WU Yan-qing[1], FAN Hai-mei[2]
(1. School of Environmental Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; 2. East China Sea Environmental Monitoring Center, State Oceanic Administration, Shanghai 200120, China)

Abstract: A new pretreatment method of the data was introduced based on the data of Environmental Monitoring Center of East China Sea Branch. With the nearest interpolation algorithm, the results could be acquired at any location, any time within research areas after eliminating periodic variation and boundary interception error. This method could be applied in the marine environment field.
Key words: nearest interpolation; high-dimension space; data pretreatment; marine environment

由于海洋面积的广阔和环境的复杂多变性, 我们无法获得像陆地上那样完整和系统的海洋环境监测资料。此外, 海洋中有些环境要素的采样和分析需要非常高昂的仪器和设备, 因此大量检测此类环境要素从成本上是不可行的, 而环境问题的复杂又使得它一般需要大量的数据来保证分析结果的可靠性[1]。所以对监测数据的充分利用就显得尤为重要。本文所提出的高维空间插值方法, 就是从原始监测数据中提取出更多信息的尝试。该方法实现了对研究区域任意时间, 地点的插值, 解决了数据的完备性问题, 在海洋环境数据预处理中有着广阔的应用前景。

1 现场调查

国家海洋局东海分局环境监测中心于 1999年到 2006年在上海海域(北纬 30°20′到 32°00′, 东经 121°01′到 124°00′)[2]进行了 28个航次的常规监测。

监测站位如图 1所示。监测采样和样品分析按照《海洋监测规范 GB 17378-1998》来进行。

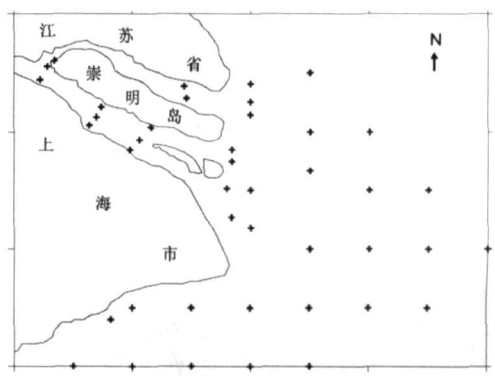

图 1 上海海域监测站点分布
Fig. 1 Sample sites in Shanghai coastal waters

常规监测分为水质监测和底泥生物监测两部分, 其中水质监测项目主要有 DO饱和 DO, pH, 浊度、悬浮物和

* 收稿日期: 2007-12-07; 修订日期: 2008-03-11
基金项目: 上海市 908专项(PJ2)
作者简介: 李 峋(1985-), 男, 山西省晋城市人, 硕士研究生, 主要从事海洋环境研究。
通讯作者: 仵彦卿, E-mail: wuyanqing@sjtu.edu.cn

油类等物理指标; COD、PO_4-P、NO_3-N、NO_2-N和 NH_4-N等营养盐指标, 以及总 Hg、Cu、Pb、Cd和 As等重金属指标。

2 数据质量分析

对于各环境要素的每个监测值, 都有一个精确到分钟的监测时间和精确到秒度的经纬度坐标相对应, 如下表所示:

表 1 监测数据格式

Tab. 1 Format of monitoring data

时间	纬度	经度	DO/mg·L^{-1}
199908020935	312957	1223000	8.10

表中 DO的监测地点为北纬 31°29′57″, 东经 122°30′00″。监测时间是 1999年 8月 2日上午 9点 35分, DO浓度为 8.10 mg/L。

在过去的文献中, 一般把一个航次的所有监测站点的监测值看作同一时间的监测值[3]。对于研究某环境要素某季节在空间上的分布, 以及该要素空间分布在不同季节的差异等问题, 简单的把一个航次的所有监测数据看作同一时间的数据, 作等值线研究其空间分布是基本可行的。

然而从严格意义上来说, 原始数据中并不存在同一时刻某环境要素在所有站点上的监测值, 也没有某环境要素在同一站点上的等时间间距序列值。前者是因为在每个航次中, 对各个站点的监测是依次进行而不是同时进行的, 后者是因为在每个站点进行监测时, 由于风浪, 水深等条件的差异无法保证每次都位于严格相同的位置上。

如果想要研究某环境要素在某一站点上的变化趋势, 就需要对该环境要素的监测值进行时间序列分析, 而时间序列分析的基本要求就是待分析的数据是等时间间隔的, 而实际采样中无法做到这一点; 此外不同航次中经纬度完全相同的监测数据很难获得, 所以无法对原始数据直接进行分析。

由于不同站点的监测值并不是在一年中的同一天或者一天中的同一时刻获得的, 所以在进行空间分析之前就必须消除监测时间不同带来的影响。此外在进行年变化和日变化分析时, 每一个航次的监测值也不能简单的看作同一时间的监测值而必须考虑其具体的监测日期和监测时刻。

由以上分析可知, 原始数据不能够直接用作此类分析, 而必须先进行数据预处理。

3 最邻近插值算法

从数学上来说, 每个监测值可以看作由时间、经度和纬度组成的三维空间中的一个点。而想要得到的是一个平面或者一条直线上点的监测值, 所以需要进行插值。

常用的插值算法有最邻近插值、线性插值、二次插值和样条插值等, 由于海洋中各监测点之间不像飞机机翼或者样条各点那样高度相关, 所以二次插值以及样条等更高次插值算法并不适用[4]; 此外由于线性插值的范围局限于由各已知监测点张成的凸线性空间之中[5], 无法获得空间中任意一点的插值结果, 因此选择最邻近插值作为海洋环境数据预处理的算法进行分析, 它也是利用其他算法进行插值计算的基础。

最邻近插值算法对已知监测值之间的关系不做任何假设, 只假设邻近空间之间的相关性。该算法认为空间中两点之间的距离越近, 这两点的属性值差异就越小, 因此就可以用距离未知点最近的已知值来对未知点的数值做出估计。使用最邻近算法, 可以获得空间中任意一点的插值结果, 但待插值点附近没有已知值时, 插值误差可能较大, 需要通过平均来降低误差。

4 周期性因素的考虑

直接在三维空间中进行最邻近插值时, 属性值的周期性波动可能会影响插值效果。例如假设某环境要素值有显著日周期变化, 那么对于同一地点的三个监测值:

(1) 1999年 10月 01日 12:00的监测值 a;

(2) 1999年 10月 02日 00:00的监测值 b;

(3) 1999年 10月 03日 12:00的监测值 c。

当 a、b、c取值如图 2所示时, 如果不考虑该环境要素的日周期变化而直接使用三维空间最临近插值算法, 则由于在时间轴上 a、b之间的距离要小于 a、c之间的距离, 所以当 a未知, 但 b、c已知时, 用 b作为 a的估计值而不是 c。然而如果考虑到日周期变化, 从 10月 01日到 10月 03日, 该环境要素值在 12:00时的差异可能要小于 12:00和 00:00之间的差异, 所以就应该用 c作为 a的估计而不是 b。

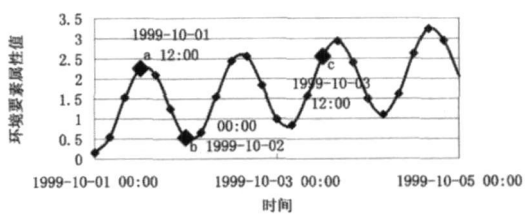

图 2 属性值周期性变化产生的插值误差

Fig. 2 Interpolation error caused by periodic variation

这种属性值在小时间尺度上的周期性波动会影响插值效果, 为了消除这种影响, 可以把小时间尺度作为独立维度提取出来参与插值过程, 这样上面三点之间的距离

就需要分别从两个维度来进行计算: a b之间相差 1 d, 但 1 d中的时刻相差 12 h, 占 1 d跨度的二分之一; 而 a c之间虽然相差 2 d, 但位于 1 d中的同一时刻。如果假设一年中 1 d的差异比 1 d中 12 h的差异要小的话, 就认为 a c距离小于 a b距离。

上面是一个假想的日周期波动的例子, 对于海洋环境中某些环境要素如温度、叶绿素含量和 DO浓度等受光照的影响较大[6,7], 所以有理由假设这些物理要素有日周期变化; 此外, 长江径流量、气温、季风和洋流等因素对各环境要素的影响也很大[8,9], 而这些因素是年周期性的, 所以还可以假设这些环境要素具有年周期变化的特性; 而由太阳, 地球, 月亮三体运动形成的复杂的潮汐变化不具有明显的和公历月份相对应的周期性和日周期性, 所以暂且不予考虑。

在要素具有年周期性和日周期性假设的基础上, 原始数据的时间维度就可以分离出两个较小的维度: 例如 1999年 8月 2日上午 9点 35分就可以分为: ① 年份(1999); ②月份(8月 2日), 需要转化为是一年中的第几天, 即年周期中的位置; ③时刻(9点 35分), 需要转化为 1 d中的第几分钟, 即日周期中的位置。

把监测值的年份数据归一化到[0, 1]之间; 月份数据换算成从这一年 1月 1日开始的天数之后除以 365, 也就实现了到[0, 1]之间的归一化; 把时刻数据换算成从每天 0点 0分开始的分钟数之后除以 24再除以 60, 也可以实现到[0, 1]之间的归一化, 然后把两个空间坐标轴经度和纬度也分别归一化到[0, 1]之间, 各要素值不作处理, 就完成了插值前的数据转换过程。

这样对于任何一个监测值 V, 都有年份值 t_1、月份值 t_2、时刻值 t_3、经度值 x、纬度值 y五个参数, 其表达式为:

$$V(t_1, t_2, t_3, x, y) \tag{1}$$

任何一个监测值 V都可以由 t_1、t_2、t_3、x、y这五个参数张成的五维空间中的一个点来表示。该空间中任意两点的距离可以定义为平方欧几里德距离, 即对于空间中任意两点:

$$V_1(t_{1_1}, t_{2_1}, t_{3_1}, x_1, y_1),$$

$$V_2(t_{1_2}, t_{2_2}, t_{3_2}, x_2, y_2)。$$

它们之间的距离为:

$$D=(t_{1_1}-t_{1_2})^2+(t_{2_1}-t_{2_2})^2+(t_{3_1}-t_{3_2})^2+(x_1-x_2)^2+(y_1-y_2)^2 \tag{2}$$

可以认为该空间中两点之间的距离越近, 属性值的差异就越小。因此可以使用离未知点最近的已知点的监测值来估计未知值, 也就是最邻近插值。

5　边界截断误差的消除

由于监测时间按照周期被分为三部分, 导致在周期边界产生了新的截断误差。例如图 3中, 对于同一地点的三个监测值:

(1) 1999年 10月 02日 01: 00的监测值 a;

(2) 1999年 10月 01日 23: 00的监测值 b;

(3) 1999年 10月 02日 12: 00的监测值 c。

按照周期截断的插值算法, 三个点在月份上的距离可以忽略, a b之间在时刻上相距 22 h, a c之间在时刻上相距 11 h, 所以 a c之间的距离小于 a b之间的距离, 可以用 c来估计 a。但实际上 a和 b之间仅相距 2 h, 应该用 b来估计 a, 人为截断周期拉长了 a b之间的距离, 产生了新的截断误差。这种误差可以通过以下方法来消除:

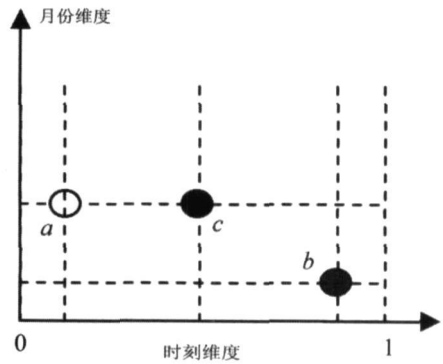

图 3　边界截断误差的产生

Fig. 3　Interpolation error caused by boundary interception

对所有监测值的时刻值进行检查, 如果时刻值 t_3大于时刻维度长度的一半 T_{50}(由于归一化到[0, 1]之间, 所以 $T_{50}=0.5$), 则将该监测值的时刻值减去 T_{50}, 其他值不变; 如果时刻值 t_3小于时刻维度长度的一半 T_{50}, 则将该监测值的时刻值加上 T_{50}, 其他值不变。这样就得到了和原数据具有相等数据量的一组新数据, 新数据和原数据在空间中的位置关系如图 4所示:

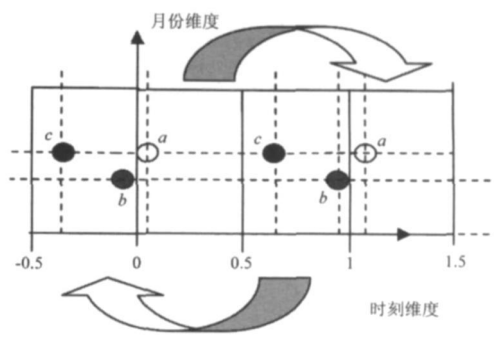

图 4　消除边界截断误差的方法

Fig. 4　Transformation to eliminate boundary interception error

可以看到, 归一化后的时刻范围从[0, 1]扩展到了[-0.5, 1.5], 如果此时再对[0, 1]范围内的点进行插值, 就可以消除由于截断周期带来的误差。例如此时对 a点的

插值结果是 b而不是 c。

日周期上截断误差的消除会导致插值数据量的倍增，如果再消除年周期上的截断误差，插值数据量就会变为原数据量的四倍，数据量的增加会影响插值计算的速度，在实际使用时需要注意。

6 距离公式的一般化

过去把一个航次某环境要素所有监测数据看作同一时间数据的处理方法实际上也相当于对时间、经度和纬度三维空间里面的属性值进行了最临近插值。经纬度的权重都取得足够大，时间轴的权重取得足够小，使得监测数据中任何时间细节上的差异都被忽略，空间中各个点的距离仅由经纬度差异的大小来决定而与时间维度无关。

而在距离公式（2）中，由于年份、月份、时刻、纬度和经度五个参数都进行了归一化，因此它们的值都位于[0 1]之间而且消除了量纲，这时各个维度的权重是相等的，属于等权重插值，而更一般的距离公式是每个维度都赋予一定的权重。因为权重的改变会对插值结果产生影响，所以可以根据实际监测结果来校正权重，使各参数的权重值达到最优，此时距离公式变为：

$$D=a*(t_1-t_2)^2+b*(\varphi_1-\varphi_2)^2+c*(\theta_1-\theta_2)^2+d*(x_1-x_2)^2+e*(y_1-y_2)^2 \quad (3)$$

7 插值时间和地点的选择

由于最临近插值在已知点附近的插值结果比较精确，所以选择在 43个监测站点坐标上进行插值，然后再利用其他插值算法绘制整个海域的空间分布图。由于监测值大都位于这 43个坐标附近，因此可以把最邻近插值的误差降低到最小限度。

在时间上，选择等时间间隔进行最邻近插值，通过平均来降低误差。插值时间根据监测时间的分布来决定：由于一天中的监测时刻大致在 06: 00到 18: 00之间均匀分布，故选择整点时刻进行插值。由于凌晨 6时之前和晚上 18时之后的监测次数较少，使得这两个时段的插值结果误差较大。因此主要对凌晨 6时到晚上 18时之间进行日变化分析，该区间之外的插值结果仅作参考。

一年中每个站点一般有 3～4次监测，监测月份和日期的差别可能较大，为均匀覆盖整个月份空间，插值时间可选择为每月的某一天（例如每月 15日）。通过试验可知，时间间隔分别为一周和一个月的插值结果平均之后在趋势变化上完全一致，只是绝对值略有差别。

8 插值结果的验证

为检验插值结果的有效性，可以对监测数据和插值结果进行对比。

利用 1999年到 2006年整个上海海域表层 DO的监测数据对 SH008站点（北纬 31. 5°，东经 122. 5°）表层 DO数据进行插值。插值地点即为站点坐标；插值时间为 1999年到 2006年间每年 2月 15日、5月 15日、8月 15日和 11月 15日中午 12: 00；插值权重比 a∶b∶c∶d∶e=10∶5∶5∶4∶4。插值结果和监测数据对比如图 5所示：

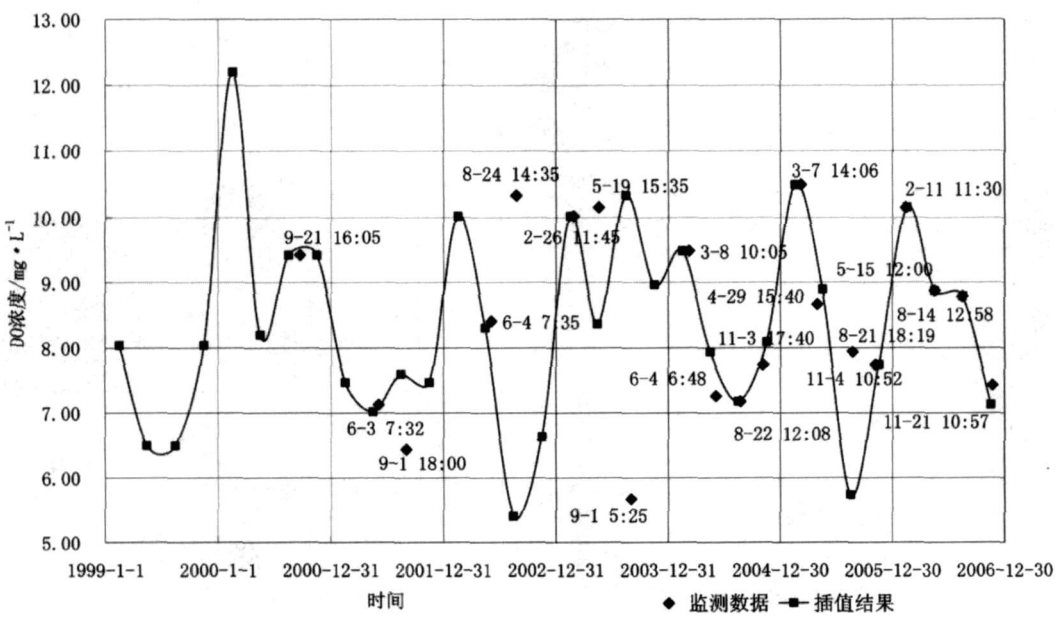

图 5 插值结果和监测数据的对比

Fig. 5 Comparison between interpolated results and monitoring data

从图 5可以看出, SH008站点从 1999年到 2006年表层 DO实际监测值较零乱, 每年的监测次数和时间不尽相同, 在 1999年甚至没有监测值。在这里, 高维空间最邻近插值算法根据邻近海域 1999年的监测值, 得到了 SH008站点 1999年的插值结果。

当监测日期和监测时刻与插值日期和插值时刻相差较大时, 插值结果和监测数据的差别较大; 反之当监测日期和监测时刻与插值日期和插值时刻相差较小时, 插值结果和监测数据的差别较小, 甚至相同。

受数据量所限, 对单一站点上单个时刻进行插值的误差可能较大, 对多个时刻插值后平均可以减小误差, 对多个站点插值后平均也可以降低误差。从不同角度研究问题时, 可以选择不同的平均方法来减小误差。

9 结　论

作为一种数据预处理方法, 高维空间最邻近插值算法实现了对分散, 零乱原始资料的系统整合, 并能够提取出新信息。因此适合于处理大区域上复杂多变的海洋数据, 实现数据的充分利用。利用该方法可以解决以下问题:

(1) 在缺乏日变化监测的情况下, 仅从监测数据中含有的时刻信息就可以进行日变化分析。

(2) 在年变化监测数据不足的情况下, 实现环境要素的年变化分析。

(3) 在每年的监测站位时间都不尽相同的情况下, 构造出各个站点上等时间间隔序列值, 对各要素的长期变化趋势进行分析。

(4) 在各监测站点无法同时进行监测的情况下, 得到同一时刻各站点的监测值, 了解其空间分布以及空间分布的变化趋势。

此外还可以进行一些更复杂, 深刻的分析, 例如可以分析不同环境要素之间的相互关系等问题。由此可以看出, 作为海洋环境数据预处理的一种基础性方法, 高维空间插值算法有着一定的应用前景。

致谢: 本文所用监测资料均由国家海洋局东海环境监测中心提供, 在此表示衷心的感谢。

参考文献:

[1] RILEY J P, SKIRROW G, 崔清晨等译. 化学海洋学 (第二卷) [M]. 北京: 海洋出版社, 1985.

[2] 刘瑞玉, 罗秉征. 三峡工程对长江口及邻近海域生态与环境的影响[J]. 海洋科学集刊, 1992, 33: 1-13.

[3] 王百顺, 刘阿成, 陈忠阳. 1984-2000年长江口海域水质重金属浓度分布变化[J]. 海洋通报, 2003, 22(2): 32-38.

[4] 李庆扬, 王能超, 易大义. 数值分析[M]. 武汉: 华中科技大学出版社, 2004.

[5] 陈宝林. 最优化理论与算法 (第二版) [M]. 北京: 清华大学出版社, 2005.

[6] 石晓勇, 陆　茸, 张传松, 等. 长江口邻近海域溶解氧分布特征及主要影响因素[J]. 中国海洋大学学报, 2006, 36(2): 287-294.

[7] 朱建荣. 长江口外海区叶绿素 a浓度分布及其动力成因分析[J]. 中国科学 D辑, 2004, 34(8): 757-762.

[8] 崔　毅, 杨琴芳, 宋云利. 夏季渤海无机磷酸盐和溶解氧分布及其相互关系[J]. 海洋环境科学, 1994, 13(4): 31-35.

[9] 杨庆霄, 董娅婕, 蒋岳文, 等. 黄海和东海海域溶解氧的分布特征[J]. 海洋环境科学, 2001, 20(3): 9-13.