第29卷第6期 计 算 力 学 学 报 Vol. 29, No. 6
2012年12月 Chinese Journal of Computational Mechanics December 2012

文章编号:1007-4708(2012)06-0879-06

基于信息扩散的海洋稀疏观测资料插值算法-椭圆模型

刘 巍[1], 张 韧*[2], 徐志升[2], 安玉柱[2], 金炜东[1]

(1.西南交通大学 信息科学与技术学院,成都 610031;
2.解放军理工大学 气象学院 军事海洋环境军队重点实验室,南京 211101)

摘 要:针对海洋科学研究中ARGO等观测资料零散、稀少等问题,提出了一种基于信息扩散思想的插值方法—信息扩散插值算法。该方法基于模糊映射思想,通过对稀少数据点信息进行模糊扩散和插值映射,进而实现有限数据点信息向其邻近区域点的概率插值;针对正态扩散函数在表现非对称结构的数据资料时存在的局限性,发展了一类非均匀信息扩散函数,建立了相应的非均匀信息扩散插值算法—椭圆模型,进行了试验对比和实际ARGO资料的插值应用。

关键词:信息扩散;插值算法;稀疏数据;椭圆模型
中图分类号:P731.1;O241.3 **文献标志码**:A

1 引 言

海洋占地球表面的绝大部分,其重要性近年来越来越得到人们关注和重视。要深入揭示和预测气候异常和海洋环境演变,需获取充分、细致和准确的海洋环境资料。然而,当前存在的最大难题之一正是海洋水文资料的极度匮乏和数据信息提取与分析处理手段的相对滞后。

ARGO计划是1998年推出的一个全球海洋联合观测调查项目,旨在快速、准确、大范围收集全球海洋温、盐剖面资料和温盐结构特征。目前,全球Argo观测网每月可提供1万多条温盐剖面实时观测数据,其数量之多是前所未有的。但ARGO资料也存在一些固有的缺陷:ARGO浮标的平均布放间隔约为300 km,浮标每10～14天完成一个海洋剖面的测量并发送数据。从区域海洋、大气研究意义上,ARGO资料在空间上是稀疏的、时间上是不连续的[1]。

插值方法是用周边资料估算和逼近缺测点数据信息的常用手段,常用的插值方法包括:Kriging插值、Lagrange插值、样条拟合插值和多项式法、有限元法、变分法以及逐步订证和最优内插方法[2,3],他们基本可满足大尺度大气、海洋数据资料的插值拟合需要。

近年来,不同方法的交叉互补,给插值方法赋予了新的内涵。如通过小波自相关函数的插值性质,可得到任意给定函数的插值小波表达[4];通过有限覆盖技术与径向点插值方法相结合,发展了有限覆盖径向点插值无网格方法[5]。

然而,常规插值方法须具备必要的站点资料、背景信息或前提条件,若观测资料过于稀少,将严重制约这些方法的准确性和可靠性。克里金插值是当前主流和行之有效的不规则数据插值技术,但常规克里金插值的变异函数为有限的确定函数,对结构复杂且稀疏的要素,难以准确有效刻画[3];最优插值和逐步订正等方法需要以要素的气候平均场作为初始"猜测场",然后用实际观测资料对该"猜测场"进行迭代和修正,然而实际海洋中即使较粗的气候背景场也很难获取[2]。因此,研究发展针对稀疏观测资料的插值方法有重要科学意义和应用价值。本文基于信息扩散思想,提出了一种针对稀疏资料的插值新方法——信息扩散插值算法。

2 信息扩散插值原理

2.1 信息扩散原理

"信息扩散"是为解决地震、风暴潮、泥石流等强灾害、小样本等自然灾害事件评估中存在的信息不完备而提出的研究思想和数学模型[6]。信息扩

收稿日期:2011-08-16;**修改稿收到日期**:2012-03-10.
基金项目:国家自然科学基金(41276088)资助项目.
作者简介:刘 巍(1982-),男,博士生;
张 韧*(1963-),男,教授,博士生导师
(E-mail: zren63@126.com).

散作为弥补信息不足而考虑优化利用样本模糊信息的一种对样本作集值化的模糊数学处理方法，它通过将单值样本转换成概率形式表达的模糊集值样本，进而对非完备样本信息进行有效处理[7]。目前，“信息扩散”仅被用于小样本事件的风险评估[8]。但其研究思想和技术途径适宜于处理稀疏资料插补和小样本信息拓展等数据不完备问题。

设 $W=\{W_1 \quad W_2\cdots W_n\}$ 是知识样本，L 是基础论域，记 W_i 的观测值为 l_i。再设 $x=\phi(l-l_i)$，则当 W 非完备时，存在函数 $\mu(x)$，使 l_i 点获得的量值为 1 的信息可按 $\mu(x)$ 的量值扩散到 l 上去。且扩散所得到的原始信息分布 $Q(l)=\sum_{j=1}^{n}\mu(x)=\sum_{j=1}^{n}\mu(\phi(l-l_i))$ 能更好地反映 W 在总体的规律，称该原理为信息扩散原理[6]。

根据信息扩散原理对母体概率密度函数的估计称为扩散估计。扩散估计的确切定义为：设 $\mu(x)$ 为定义在 $(-\infty,+\infty)$ 上的一个波雷尔可测函数，$d>0$ 为常数，$x=(l-l_i)/d$，则称

$$\hat{f}(l)=\frac{1}{nd_i}\sum_{i=1}^{n}\mu[(l-l_i)/d] \tag{1}$$

为母体概率密度函数 $f(l)$ 的一个扩散估计。式中 $\mu(x)$ 为扩散函数，d 为窗宽。

2.2 信息扩散插值思想

本文将“信息扩散”思想移植运用于稀疏数据的拟合插值，提出了一种适宜于处理稀疏资料和小样本数据的插值新方法——信息扩散插值算法。

2.2.1 “输入-输出”模糊映射关系

对于一个“输入-输出”系统，记 Ω 为母体，x 为“输入”变量，y 为“输出”变量，X 为“输入”集合，Y 为“输出”集合。即

$$x\in X,\ y\in Y,\ \Omega=X\times Y$$

令 $f(x,y)$ 为母体 Ω 的概率密度函数。则 y 满足 $x=u$ 的条件概率密度为

$$f_{Y|X}(y|u)=f(u,y)\Big/\int_{v\in Y}f(u,v)\mathrm{d}v \tag{2}$$

上式可理解为：从母体 Ω 中随机抽取一个样本，该样本“输入”为 u 时“输出”为 y 的可能性。亦即，对于固定的“输入-输出”系统，给定其“输入”，对应的“输出”具有一定的概率分布形式。

基于模糊集合思想，定义 Ω“输入-输出”系统为在给定的“输入”下的“输出”为一个模糊集 $\widetilde{B}$。模糊集 $\widetilde{B}$ 的隶属度函数与“输出”的概率密度函数相对应（概率密度值大对应隶属度高，反之则低），且概率密度函数的归一化结果即为隶属度函数，即模糊集 $\widetilde{B}$ 的隶属度函数为

$$\begin{aligned}\mu_{\widetilde{B}}(y)&=\frac{f(u,y)\Big/\int_{v\in Y}f(u,v)\mathrm{d}v}{\max\limits_{y\in Y}\left\{f(u,y)\Big/\int_{v\in Y}f(u,v)\mathrm{d}v\right\}}\\&=f(u,y)/\max_{y\in Y}\{f(u,y)\}\end{aligned} \tag{3}$$

这样，在该“输入-输出”系统中，任意给定一个“输入”$x(x\in X)$，其所有可能的“输出”可以用模糊集合 $\widetilde{B}$ 来表示，这就是“输入-输出”系统中“输入”、“输出”间的模糊映射关系。

但母体 Ω 的概率密度函数 $f(x,y)$ 并不能直接获得，一般需用大量样本的统计结果来进行估计。但在数据资料稀缺的情况下，可以从获取的少量样本数据出发，通过信息扩散得到其总体概率密度分布的一个近似估计。

2.2.2 信息扩散函数估计

设 S 为获取的 Ω“输入-输出”系统的一组小样本，记为 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$。

由于样本数量小，不满足常规统计回归的样本长度要求和信度检验，不能用常规统计分析方法来计算概率密度函数。同时，用常规统计方法得到的非完备小样本的概率密度函数亦无法反映母体 Ω 的概率密度分布情况。

信息扩散中将小样本 S 的数据看成散布在“输入-输出”论域空间 $X\times Y$ 上的“信息注入点”，通过集值化处理，将其每一个样本数据扩展为周围多个样本点模糊集的表示形式，即每一个样本点将充当“周围未出现样本点的代表”。由于“周围”的边界不清楚、模糊和弹性，所以每一个样本点所提供的包括周围影响在内的信息总体是一个模糊信息。针对样本点 (x_i,y_i)“周围”边界的不清楚和模糊性特征，在输入空间 X 和输出空间 Y 中分别引入“监控点”$u_j(j=1,2,\cdots,s)$ 和 $v_k(k=1,2,\cdots,t)$，它们是“输入”、“输出”空间上的标准化离散点。将“输入”、“输出”监控点集合分别记为 $U=\{u_1,u_2,\cdots,u_s\}$，$V=\{v_1,v_2,\cdots,v_t\}$，这样监控点空间 $U\times V$ 就构成了分布在“输入-输出”空间上的网格。将“信息注入点”的信息通过某种形式（信息扩散公式）合理、有效地扩散到整个监控点空间上，即可实现非完备样本数据信息的拓展和挖掘。

记 A=Ω×U×V，定义论域 A 到区间[0，1]的一个映射 μ:A→[0,1] 使得

$$((x,y),u,v)\in A\rightarrow\mu_{u_jv_k}(x_i,y_i)\in[0,1]$$

式中 $(x,y)\in\Omega$，$u\in U$，$v\in V$，$j=1,2,\cdots,s$，$k=1,2,\cdots,t$，$i=1,2,\cdots,n$，称 $\mu_{u_jv_k}(x_i,y_i)$ 为信息扩散函数。

令 $q_{u_jv_k}=\sum_{i=1}^{n}\mu_{u_jv_k}(x_i,y_i)$，简记为 q_{jk}，$t=\sum_{j=1}^{s}\sum_{k=1}^{t}q_{jk}$，通过信息扩散，该 Ω“输入-输出”系统在点 (u_j,v_k) 处的概率密度函数估计可表示为

$$\hat{f}(u_j,v_k)=q_{jk}/t \tag{4}$$

2.2.3　信息扩散插值映射

将概率密度函数的扩散估计 $\hat{f}(u_j,v_k)$ 代入“输入-输出”映射关系式(3)，当输入为 u 时，输出即为

$$\widetilde{B}=\int_{y\in Y}\mu_{\widetilde{B}}(y)/y=\int_{y\in Y}\hat{f}(u,y)/\max_{y\in Y}\{\hat{f}(u,y)\}\Big/y \tag{5}$$

这样，根据信息扩散原理，从可获取的少量样本信息出发，建立从“输入”到“输出”的模糊映射关系，通过对“输出”进行去模糊化处理即得到了插值结果。

3　非均匀信息扩散插值-椭圆模型

信息扩散的核心在于构建一个准确、合理的扩散函数。黄崇福模仿分子扩散，推导得到了正态信息扩散函数[6,7]。正态信息扩散函数表现的是以资料点为中心，向周边均匀扩散并随空间距离逐渐递减的方式，它反映的是一种均匀扩散过程，但实际数据样本(如大气、海洋资料)之间可能存在复杂的结构和显著的差异，数据结构更多表现出非对称的特征。因此，正态扩散模型只是实际数据结构中较为理想的特例。客观、合理地描述和刻画实际样本数据中更广义的非正态、非均匀的数据结构，必须考虑更逼近实际的非均匀信息扩散模型。

为此，本文提出了一种“椭圆式”非均匀信息扩散新思想，即认为数据点信息不是按照正态分布函数以“圆对称”的形式向四周均匀扩散，而是以一种“椭圆”形式非均匀扩散，即沿某方向可能扩散得快些(定义为椭圆长轴)，与其正交的方向可能扩散得慢些(定义为椭圆短轴)。这样就将均匀的“圆形”正态信息扩散拓展为“椭圆”形式的非均匀信息扩散，如图 1 所示。

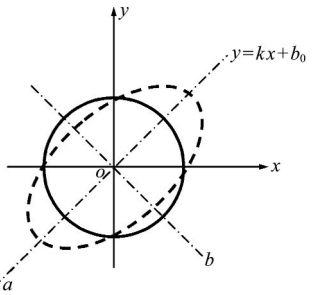


图 1　正态(实线圆)和非均匀(虚线椭圆)式信息扩散示意图

Fig. 1　Information diffusion frame，normal(solid line round) and unsymmetrical(dashed line ellipse)

3.1　二维“椭圆式”扩散函数

在二维数据结构中，设定扩散快的方向与椭圆的长轴(直线 a)对应，扩散慢的方向则与椭圆的短轴(直线 b)对应。这样，信息扩散函数就变为

$$\mu=\frac{1}{2\pi h_xh_y}\exp\left\{-\frac{1}{k^2+1}\left[\frac{1}{\lambda}\left(\frac{x}{\sqrt{2}h_x}+k\frac{y}{\sqrt{2}h_y}\right)^2+\left(k\frac{x}{\sqrt{2}h_x}-\frac{y}{\sqrt{2}h_y}\right)^2\right]\right\} \tag{6}$$

式中 k 为椭圆长轴的斜率，定义为“旋转系数”，λ 为椭圆长轴与短轴之比的平方，定义为“伸缩系数”，称该扩散函数为“椭圆式”非均匀信息扩散函数。

3.2　k 与 λ 参数的确定

在二维“椭圆式”非均匀信息扩散函数中，引入了两个重要参数：旋转系数 k 和伸缩系数 λ。

其中旋转系数 k 是椭圆长轴的斜率，它与样本在 xOy 平面上的分布有关，沿椭圆长轴方向扩散最快；因此，样本点沿长轴所在直线周围分布的可能性也就最大，可以认为各样本点 (x_i,y_i) 到该直线 $y=kx+b_0$ 的距离平方和最小，即

$$Q=\sum_{i=1}^{n}d_i^2=\sum_{i=1}^{n}\frac{(kx_i+b_0-y_i)^2}{k^2+1}=\min \tag{7}$$

故有：

$$\begin{cases}\dfrac{\partial Q}{\partial b_0}=\sum\limits_{i=1}^{n}2(kx_i+b_0-y_i)/(k^2+1)=0\\ \dfrac{\partial Q}{\partial k}=\sum\limits_{i=1}^{n}[2(kx_i+b_0-y_i)(k^2+1)-\\ \qquad 2k(kx_i+b_0-y_i)^2]\Big/(k^2+1)^2=0\end{cases} \tag{8}$$

化简得到：

$$\begin{cases}b_0=\dfrac{1}{n}\left(\sum\limits_{i=1}^{n}y_i-k\sum\limits_{i=1}^{n}x_i\right)\\ k^2\cdot\sum\hat{x}_i\hat{y}_i+k\cdot\sum(\hat{x}_i^2-\hat{y}_i^2)-\sum\hat{x}_i\hat{y}_i=0\end{cases} \tag{9}$$

式中 $\hat{x}_i=x_i-\sum x_i/n$，$\hat{y}_i=y_i-\sum y_i/n$。进而可以求

解出 b_0 和 k。

伸缩系数 λ 反应了椭圆的“胖瘦”程度，可用样本点到短轴所在直线距离的平均值（或最大值）与到长轴所在直线距离的平均值（或最大值）之比来表示。伸缩系数 λ 可看成是各样本点到直线 b 距离的平均值（或最大值）与到直线 a 距离的平均值（或最大值）之比（图 1）。

3.3 多维“椭圆式”扩散函数

多维（维数为 $m+1$）正态信息扩散函数的无量纲形式为

$$\mu=\frac{1}{(2\pi)^{(m+1)/2}h_y\prod_{i=1}^{m}h_{x_i}}\exp\left[-\left(\sum_{i=1}^{m}x_i'^2+y'^2\right)\right] \tag{10}$$

将该正态“圆形”均匀扩散函数拓展为“椭圆”非均匀扩散函数，即引入伸缩系数，得到：

$$\mu=\frac{1}{(2\pi)^{(m+1)/2}h_y\prod_{i=1}^{m}h_{x_i}}\exp\left[-\left(\sum_{i=1}^{m}\frac{x_i'^2}{\lambda_i}+y'^2\right)\right] \tag{11}$$

再在该非均匀扩散函数的指数项 $-\sum_{i=1}^{m}\frac{x_i'^2}{\lambda_i}-y'^2$ 中考虑坐标旋转（引入旋转系数）因素，使得扩散快、慢的方向与其相应投影平面上的长轴、短轴对应。即分别对 $x_i'(i=1,\cdots,m)$ 依次作如下坐标变换：

$$\begin{cases}x_i'=x_i'\cos\theta_i+y'\sin\theta_i\\ y'=-x_i'\sin\theta_i+y'\cos\theta_i\end{cases} \tag{12}$$

得到：

$$\mu=\frac{1}{(2\pi)^{(m+1)/2}h_y\prod_{i=1}^{m}h_{x_i}}\exp\left\{-\sum_{i=1}^{m}\frac{1}{\lambda_i}\left[x_i'\cos\theta_i-\sum_{j=i+1}^{m}\left(x_j'\sin\theta_j\sin\theta_i\prod_{k=i+1}^{j-1}\cos\theta_k\right)+y'\sin\theta_i\prod_{j=i+1}^{m}\cos\theta_j\right]^2-\left[-\sum_{i=1}^{m}\left(x_i'\sin\theta_i\prod_{j=1}^{i-1}\cos\theta_j\right)+y'\prod_{i=1}^{m}\cos\theta_i\right]^2\right\} \tag{13}$$

式中 $x_i'=x_i/\sqrt{2}\,h_{x_i}$，$y'=y/\sqrt{2}\,h_y$，$\cos\theta_i=1/\sqrt{k_i^2+1}$，$\sin\theta_i=k_i/\sqrt{k_i^2+1}$，$\lambda_i$ 为伸缩系数，k_i 为旋转系数。这样即得到了多维（维数为 $m+1$）的非均匀信息扩散函数。

如果输入项 x_i' 和 $x_j'(i\neq j)$ 之间有相关性，可对上式再作如下变换：

$$\begin{cases}x_i'=x_i'\cos\theta_{ij}+x_j'\sin\theta_{ij}\\ x_j'=-x_i'\sin\theta_{ij}+x_j'\cos\theta_{ij}\end{cases} \tag{14}$$

从而进一步地在扩散函数中反应出输入因子之间的相互关系。

4 算法试验

为检验信息扩散插值算法的有效性和可靠性，采用美国大气环境预报中心（NCEP）/美国大气环境研究中心（NCAR）提供的海面温度（SST）再分析资料进行插值试验和不同算法的对比分析。资料时间：2009 年逐月平均；资料范围：100°E～250°E，0°～60°N 海区；资料分辨率：2°×2°，共有 75（纬向）×30（经向）＝2250 个网格数据点，扣除陆地后有约近 2000 个格点数据。

试验一 资料：从上述 2000 个数据点中随机抽取 1%样本（23 个）作为“观测”资料（其余格点视为资料缺测），分别用 Kriging 模型和正态扩散以及非均匀“椭圆”扩散模型进行数据插补试验和对比分析。

试验目的：检验不同方法用该 1%“观测”资料插值逼近原有海温场的准确率和可靠性。

图 2 海温插值场与实际场的相关系数和误差均方差对比结果表明，信息扩散插值较之 Kriging 方法更为准确，尤以非均匀“椭圆”模型的相关系数最大、误差均方差最小，效果最佳。

试验二 资料、区域、时间同于实验一，但插值样本点增至 10%，即从数据点中随机抽取 10%的样本（230）作为“观测”资料（其余格点视为资料缺测），用 Kriging 模型和正态扩散以及非均匀“椭圆”扩散模型进行数据插补试验和对比分析，检验不同方法用该 10%的“观测”点资料插值出逼近原有海温场的准确率和可靠性。

图 3 海温插值场与实际场的相关系数和误差均方差对比结果表明，不同方法之间插值效果比较接近，差别较小（相关系数最大相差 0.002；均方差最大相差 0.2，较之试验一小了约 1 个量级）。相比而言，Kriging 方法较之信息扩散插值更为有效，表明 Kriging 插值作为经典成熟的插值算法在常规插值分析中的可靠性和有效性。

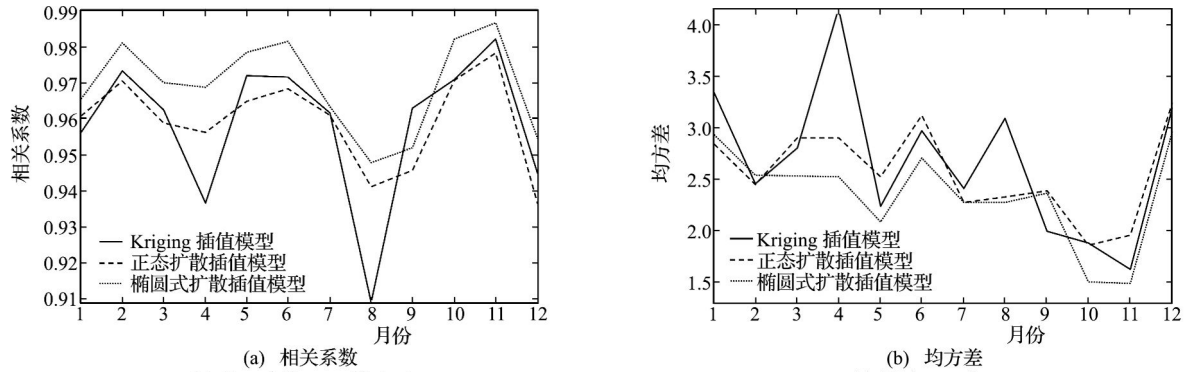
试验一与试验二的对比结果表明：本文提出的基于小样本信息扩散的插值方法（尤其是非均匀“椭圆”模型），在处理数据样本极为稀疏的海洋要素插值中较 Kriging 等插值方法表现出较为明显的优势，而随着样本信息增加其优势逐渐减退（2%～

5%的样本试验中仍保持一定程度的优势，图略）。因此，非均匀信息扩散方法更适宜分析处理数据资料极为稀疏的信息不完备的插值计算。

算法应用：将信息扩散插值方法应用于实际ARGO散点资料的要素场标准化分析处理。资料：2009年1月1日Argo浮标10 m深层海温观测资料；范围：100°E～250°E，0°N～60°N海区。

图4为Argo浮标资料（图中小圆点为浮标位置，约80余个）和均匀信息扩散和非均匀"椭圆"模型插值得到的1°×1°的网格化温度场资料。该区域的数据网格点约为4000余个（扣除陆地），因此所得结果相当于2% Argo稀疏样本插值所得。

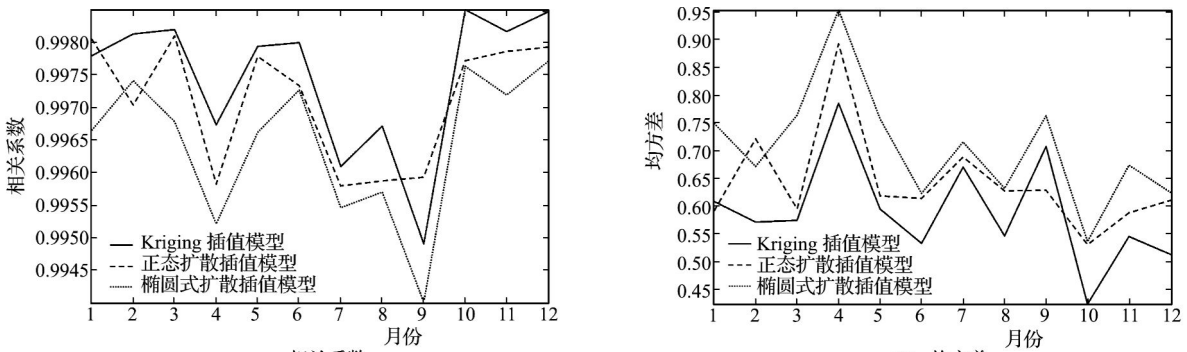
由于实时观测的海洋网格化资料很难获取，因此仅可对图4结果定性描述。总的来看，两种插值结果均表现出了冬季太平洋海区海温场的基本结构特征，如西太平洋暖池区和中高纬较密的海温梯度。但相比而言，椭圆模型插值结果的细节刻画更丰富一些。



(a) 相关系数　(a) Correlation coefficient　(b) 均方差　(b) Variance of mean

图2　不同算法的插补效果比较（1%样本）

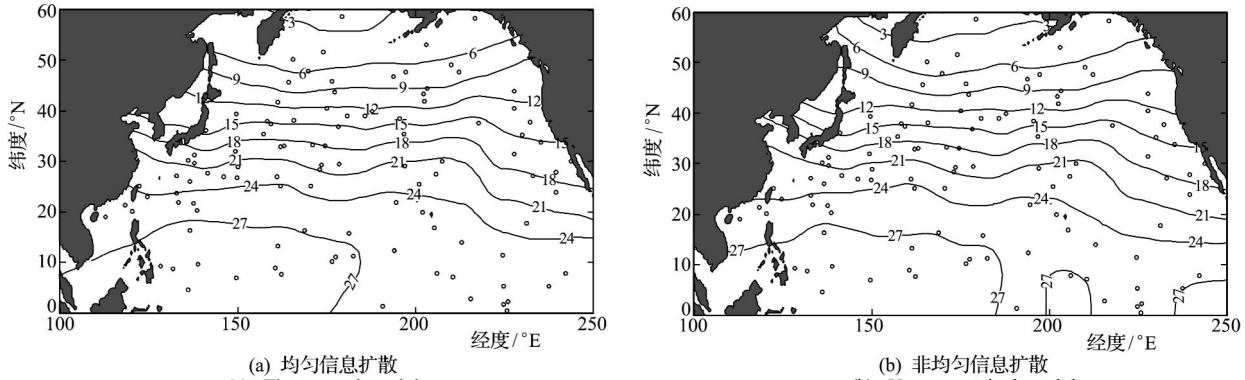
Fig. 2　The contrast between different interpolation techniques(1% sample only)



(a) 相关系数　(a) Correlation coefficient　(b) 均方差　(b) Variance of mean

图3　不同算法的插补效果比较（10%样本）

Fig. 3　The contrast between different interpolation techniques(10% sample)



(a) 均匀信息扩散　(a) The normal model　(b) 非均匀信息扩散　(b) Unsymmetrical model

图4　Argo观测资料的海温场格点插值场

Fig. 4　The sea temperature information diffusion interpolation gridding field based on Argo observational data

5 结 语

针对大气、海洋科学中广泛存在的观测资料零散、稀少等问题，提出了一种新的数据插值方法——信息扩散插值算法。该方法基于模糊集思想，通过构建逼近目标资料分布结构的扩散函数来对有限的稀少数据点信息进行模糊集扩散和插值映射，实现有限数据点信息向其邻近区域点的预估插值。针对正态扩散函数在表现非对称结构的数据资料时存在的局限性，发展了一类非均匀信息扩散函数，建立了相应的非均匀信息扩散插值算法-椭圆模型。试验对比分析表明，信息扩散插值方法为自然科学领域中实际存在的小样本和稀疏数据的分析处理作了有益的探索，提供了可资参考的方法和途径。

参考文献(References)：

[1] 许建平. ARGO 全球海洋观测探秘[M]. 北京：海洋出版社，2002. (XU Jian-ping. *ARGO Global Oceanic Observation System Exploring*[M]. Beijing: Ocean Press, 2002. (in Chinese))

[2] 封国林，董文杰. 观测数据非线性时空分布理论和方法[M]. 北京：气象出版社，2006. (FENG Guo-lin, DONG Wen-jie. *The Theory and Method for Dealing with Nonlinear Space-time Pattern of Observational Data*[M]. Beijing: Meteorological Press, 2006. (in Chinese))

[3] 张 韧，万齐林. 地学资料中的散乱数据优化与缺损信息恢复[J]. 数据采集与处理，2006，**21**(2)：209-216. (ZHANG Ren, WAN Qi-lin. Scattered data optimization and imperfect information recovery in geoscience [J]. *Journal of Data Acquisition and Processing*, 2006, **21**(2): 209-216. (in Chinese))

[4] 樊 成，栾茂田，黎 勇. 有限覆盖径向点插值方法理论及其应用[J]. 计算力学学报，2007，**24**(3)：306-311. (FAN Cheng, LUAN Mao-tian, LI Yong. The radial point-interpolation procedure based on finite covers and its applications[J]. *Chinese Journal of Computational Mechanics*, 2007, **24**(3): 306-311. (in Chinese))

[5] 宗 智，赵 勇. 小波插值方法自适应求解时间进化微分方程[J]. 计算力学学报，2010，**27**(1)：65-69. (ZONG Zhi, ZHAO Yong. Numerical solution for differential evolutional equation using adaptive interpolation wavelet method[J]. *Chinese Journal of Computational Mechanics*, 2010, **27**(1): 65-69. (in Chinese))

[6] 黄崇福. 自然灾害风险评价-理论与实践[M]. 北京：科学出版社，2006. (HUANG Chong-fu. *Natural Disaster Risk Assessment-Theory and Practice*[M]. Beijing: Science Press, 2006. (in Chinese))

[7] Huang C F. Information diffusion technique and small sample problem[J]. *Information Technology and Decision Making*, 2002, **1**(2): 229-249.

[8] 张继权，李 宁. 主要气象灾害风险评估与管理数量化应用[M]. 北京：北京师范大学出版社，2007. (ZHANG Ji-quan, LI Ning. *Quantitative Methods and Application of Risk Assessment and Management on Main Meteorology Disasters*[M]. Beijing: Beijing Normal University Press, 2007. (in Chinese))

Ellipse model, an algorithm for sparse data interpolation based on information diffusion

LIU Wei[1], ZHANG Ren[*2], XU Zhi-sheng[2], AN Yu-zhu[2], JIN Wei-dong[1]

(1. School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China;
2. Institute of Meteorology, PLA University of Science and Technology, Nanjing 211101, China)

Abstract: Aiming at the difficulties of scattered and sparse observational data in ocean science, a new interpolation algorithm based on information diffusion idea was presented in this paper. By fuzzy mapping route, the sparse data samples was diffused and mapped into corresponding fuzzy sets in the form of probability in the interpolation ellipse model. For avoiding the shortcoming of data asymmetrical structure in normal diffusion function, a kind of unsymmetrical information diffusion function was developed, and the corresponding unsymmetrical information diffusion algorithm-ellipse model was established. By making the interpolation experiments and contrast for ARGO sea surface temperature data, the rationality and validity of the ellipse-model were validated.

Key words: information diffusion; interpolation algorithm; sparse data; ellipse model