第 20 卷第 4 期
2006 年 7 月

常熟理工学院学报
Journal of Changshu Institute of Technology

Vol. 20 No. 4
July. 2006

浅析 2005 数学建模竞赛 C 题*

金　健,朱惠健

(常熟理工学院 数学系,江苏 常熟　215500)

摘　要: 探讨了 2005 数学建模竞赛 C 题"雨量预报方法的评价"的命题意图,结合我校学生的竞赛论文分析了某些参赛队解题中的典型错误,并提供了用散乱数据的曲面拟合求解本题的一些关键要点。

关键词: Shepard 插值;散乱数据插值;曲面拟合

中图分类号: O141.4　　**文献标识码**: A　　**文章编号**: 1008 - 2794(2006)04 - 0041 - 05

散乱数据插值(scattered data interpolation)主要研究根据给定散乱数据点构造光滑曲面的理论与方法,其历史可追溯到二十世纪二十年代,目前散乱数据插值技术已广泛应用于各类科学研究和工程技术中,如气象、勘探、医学、环保、可视化以及测量造型等,散乱数据插值问题的提法是:给定数据 $(X_i, f_i) \in R^n \otimes R, i = 1,2,\cdots,N$,要求构造函数 $G = G(X)$ 插值给定的函数值,即 $G(X_i) = f_i, i = 1,2,\cdots,N$,插值方法的优劣可以从精确性、视觉效果、对参数的灵敏性、执行时间、存储量要求和在计算机上实现的简易性等方面进行比较[1],使用何种插值方法需视问题的实际情况而定。

1 "雨量预报评价"的插值模型

1.1　题目

雨量预报对农业生产和城市工作和生活有重要作用,但准确、及时地对雨量作出预报是一个十分困难的问题,广受世界各国关注。我国某地气象台和气象研究所正在研究 6 小时雨量预报方法,即每天晚上 20 点预报从 21 点开始的 4 个时段(21 点至次日 3 点,次日 3 点至 9 点,9 点至 15 点,15 点至 21 点)在某些位置的雨量,这些位置位于东经 120 度、北纬 32 度附近的 53 ×47 的等距网格点上。同时设立 91 个观测站点实测这些时段的实际雨量,由于各种条件的限制,站点的设置是不均匀的。气象部门希望建立一种科学评价预报方法好坏的数学模型与方法。气象部门提供了 41 天的用两种不同方法的预报数据和相应的实测数据。预报数据在文件夹 FORECAST 中,实测数据在文件夹 MEASURING 中,雨量用毫米做单位,小于 0.1 毫米视为无雨。

(1) 请建立数学模型来评价两种 6 小时雨量预报方法的准确性;

(2) 气象部门将 6 小时降雨量分为 6 等:0.1 - 2.5 毫米为小雨,2.6 - 6 毫米为中雨,6.1 - 12 毫米为大雨,12.1 - 25 毫米为暴雨,25.1 - 60 毫米为大暴雨,大于 60.1 毫米为特大暴雨。若按此分级向公众预报,如何在评价方法中考虑公众的感受?

1.2　解题分析

* 收稿日期:2006 - 02 - 22
作者简介:金　健(1964 —),男,江苏常熟人,常熟理工学院数学系讲师。

对于问题一,首先要找出预报误差,并由预报误差采取各种合理的定义给出评价准则函数,从而评判两种预报方法的优劣。由于预报值与实测值分布在不同经纬度的地点上,为了得到预报误差,需要做插值,可以将 91 个观测点上的实测数据进行插值,得到 53 ×47 个网格点的实测值,并与 53 ×47 个网格点上的预报值比较误差;也可以将 53 ×47 个网格点上的预报数据进行插值,得到 91 个观测点上的预报值,并与 91 个观测点上的实测值比较误差。前一种方案中,注意到观测站点的分布是不均匀的,插值方法可以选用散乱数据的曲面拟合方法,如径向基插值,Shepard 插值[2];后一种方案中,网格点的分布是均匀的,插值方法可以选用线性插值,样条插值等,且可直接调用 matlab 的插值命令。在直角坐标系中插值必须先将各站点的经纬度数据转换为直角坐标。通过插值得到误差后,可建立相对误差平方和或绝对误差平方和等评价准则函数评价两种预报方法的优劣,采用相对误差平方和准则时要注意处理分母为零的问题。另外本题的一个特点是数据量极大,雨量值个数为 2 ×41 ×4 ×53 ×47 + 41 ×4 ×91≈84(万个),分布在 200 多个文件里。

对于问题二,可将预报值与实测值先按分级预报标准换成等级,再将预报误差改成等级差。考虑到采用等级预报中公众的心理因素,譬如预报方法甲,5 次预报中 4 次无等级差,但 1 次就误差了 3 个等级;预报方法乙,5 次预报中 2 次无等级差,3 次都误差 1 个等级,两者 5 次预报虽然都误差了 3 个等级,但由于相邻等级雨量有部分值较为接近,如实际雨量为 2.6 毫米,预报等级为小雨,此时虽然误差一个等级,但也会有人觉得预报是正确的,故公众会觉得预报方法甲比预报方法乙差。由此可见,建立评价准则函数时不光要考虑预报等级差,还要对预报等级差加不同的权重,使得预报与实测的级差越大,评价越低。若将有雨报成无雨或将无雨报成有雨,也应给予"惩罚"。

"雨量预报方法的评价"题考核了散乱数据的曲面拟合方法,拟合曲面不光是 CAD/ CAM 技术的迫切需求,在地球科学、气象、航空航天、国防等关系国计民生的行业也有着及其重要的地位,散乱数据的拟合和插值有许多种不同方法,但由于应用问题的千差万别,数据量大小不同,对连续性的要求也不同等等,没有一种算法适用于所有场合。[3]考虑到本题的实际情况,可以选用径向基插值、Shepard 插值等[4],下面给出雨量预报评价的 Shepard 插值模型,matlab 计算程序及计算结论。

1.3 雨量预报评价的 Shepard 插值模型

1.3.1 符号说明

i:网格点序号$(i = 1,2,\cdots,2491)$ j:观测站点序号$(j = 1,2,\cdots,91)$ k:预报或实测的次数$(k = 1,2,\cdots,164)$ f_j^k:第 j 个观测站点的雨量的第 k 次实测值 (θ_i,φ_i):第 i 个站点的经度,纬度 d_{ij}:第 j 个观测站点到第 i 个网格点的距离 g_i^k:第 i 个网格点第 k 次预报雨量的插值值 h_i^k:第 i 个网格点第 k 次预报雨量值 t_i^k:第 i 个网格点第 k 次预报雨量的插值等级值 l_i^k:第 i 个网格点第 k 次预报雨量等级值 w^k:第 k 次预报绝对误差的算术平均数 w:164 次全部预报 w^k 总和 r:地球半径

1.3.2 模型建立

Shepard 插值是气象学家 Shepard 最早提出的应用于气象研究的方法,该方法的描述如下:

令 ρ 为平面上某一度量,通常取为距离度量。对定点 (x,y),令 $r_i = \rho((x,y),(x_i,y_i))$ $i = 1,2,\cdots,N$ 设 μ 为一个正实数,对散乱数据点 (x_i,y_i,f_i),$i = 1,2,\cdots,N$,则拟合曲面 $z = f(x,y)$ 表示成下列插值公式:

$$z = f(x,y) = \begin{cases} \dfrac{\sum_{i=1}^{N} \frac{f_i}{r_i^{\mu}}}{\sum_{i=1}^{N} \frac{1}{r_i^{\mu}}} & r_i \neq 0 \\ f_i & r_i = 0 \end{cases} \qquad (1)$$

这是一个关于 (x_i,y_i),$i = 1,2,\cdots,N$ 的全局插值公式,当 (x,y) 是非插值点时,$f(x,y)$ 取所有函数值 f_i 的权平均;权因子 $(1/r_i^{\mu})$ 与 (x,y) 有关,(1) 式称为 Shepard 公式。Shepard 方法可以处理观测数据的数目很大,对于有峰值的曲面,有良好的逼近效果。在下面的模型中,取 $\mu = 2$,也即通常所说的反距离平方插值。

为了方便,将网格点依次编号 1,2,…2491,将观测站点依次编号 1,2,…91,将 41 天每天四次的预报依次编号 1,2,…164, 文件夹 FORECAST 中文件 xx1_dis1,表示第一种预报方法在第 x 天第 x_1 时段预测值,文件 xx1_dis2,表示第二种预报方法在第 x 天第 x_1 时段预测值,文件夹 MEASURING 中文件 x,表示第 x 天观测站点实测雨量值,取地球半径 $r = 6371$km.

建立空间直角坐标系,以地心为 O 点,X 轴正向指向东经 0 度,Y 轴正向指向东经 90 度,Z 轴正向指向北极,则球面上经纬度为 (θ_i,φ_i) 点的直角坐标为 $(r\cos\varphi_i\cos\theta_i, r\cos\varphi_i\sin\theta_i, r\sin\varphi_i)$,由此得观测站点 (θ_j,φ_j) 与网格点 (θ_i,φ_i) 的距离平方公式为

$$d_{ij}^2 = r^2[(\cos\varphi_i\cos\theta_i - \cos\varphi_j\cos\theta_j)^2 + (\cos\varphi_i\sin\theta_i - \cos\varphi_j\sin\theta_j)^2 + (\sin\varphi_i - \sin\varphi_j)^2] \tag{2}$$

故第 i 个网格点第 k 次预报值为

$$g_i^k = \begin{cases} f_j^k, & d_{ij} = 0 \\ \sum_{j=1}^{91} f_j^k / d_{ij}^2 \Big/ \sum_{j=1}^{91} \frac{1}{d_{ij}^2}, & d_{ij} \neq 0 \end{cases} \tag{3}$$

第 i 个网格点第 k 次预报的绝对误差为 $|g_i^k - h_i^k|$,第 k 次预报绝对误差平均值

$$w^k = \frac{1}{53\times 47}\sum_{i=1}^{53\times 47}|g_i^k - h_i^k| \tag{4}$$

全部 164 次预报误差

$$w = \sum_{k=1}^{164} w^k \tag{5}$$

利用公式(2)至公式(5)可分别计算出两种预报方法的 w 值,w 值小的预报方法优于 w 值大的预报方法。

在问题二中,先将网格点上的预报值 h_i^k $(i = 1,2,\cdots 2491;\ k = 1,2,\cdots 164)$ 及插值值 g_i^k 按照题中分级标准转化成等级值 l_i^k, t_i^k $(l_i^k, t_i^k = 0,1,2,3,4,5,6)$,记第 i 个网格点第 k 次预报等级的绝对误差为 x,即 $x = |t_i^k - l_i^k|$,为了得到公众"敏感度"的量化指标 $h(x)$,注意到人们对误报等级差得多的敏感度远远大于误报等级差得少的敏感度,为此,可以取近似的偏大型柯西分布隶属函数如下:

$$h(x) = \begin{cases} [1 + \alpha(x - \beta)^{-2}]^{-1}, & 0 \le x \le 4 \\ a\ln x + b, & 4 \le x \le 6 \end{cases}$$

其中 α,β,a,b 为待定常数。实际上,当"很敏感"时,则敏感度的量化值为 1,即 $h(6) = 1$;当"比较敏感"时,则敏感度的量化值为 0.8,即 $h(4) = 0.8$;当"很不敏感"时,则敏感度的量化值为 0.01,即 $h(0) = 0.01$。于是,利用 matlab 的非线性最小二乘估计[5]命令可以确定出 $\alpha = 3.6264$,$\beta = 0.1914$,$a = 0.4932$,$b = 0.1163$。故可以得到相应的隶属函数。经计算 $h(0) = 0.01$,$h(1) = 0.15276$,$h(2) = 0.47424$,$h(3) = 0.68506$,$h(4) = 0.8$,$h(5) = 0.9101$,$h(6) = 1$,即公众对各级降雨量预报等级差的感受量化值为{0.01,0.15276,0.47424,0.68506,0.8,0.9101,1}。结合公众的感受第 i 个网格点第 k 次预报的误差为 $h(|t_i^k - l_i^k|) \times |t_i^k - l_i^k|$,第 k 次预报的平均误差为 $\frac{1}{2491}\times\sum_{i=1}^{2491} h(|t_i^k - l_i^k|) \times |t_i^k - l_i^k|$,全部 164 次预报的总误差为 $\sum_{k=1}^{164}\frac{1}{2491}\times\sum_{i=1}^{2491} h(|t_i^k - l_i^k|) \times |t_i^k - l_i^k|$,对两种预报方法而言,以总误差值小的为佳。模型的求解利用 matlab6.5 数学软件编程计算,程序如下:

1.3.3　matlab 程序及计算结果

```
P = zeros(53,47);dd = 0;ff = 0;sigma1 = 0;sigma2 = 0;ab = 0;ac = 0;
load FORECAST\lat.dat;load FORECAST\lon.dat;load MEASURING\020618.SIX;
weidu = lat;jingdu = lon;wp = X020618(:,2);jp = X020618(:,3);
for m = 1:41    m, x = num2str(m);y = strcat('MEASURING\',x);z = textread(y);
        for t = 1:4
x1 = num2str(t);y1 = strcat('FORECAST\',x,x1,'-dis1');y2 = strcat('FORECAST\',x,x1,'-dis2');
    z1 = textread(y1);A1 = z1(:,1:47);z2 = textread(y2);A2 = z2(:,1:47);p = z(:,t+3);
for i = 1:53    for j = 1:47
```

```
d1 = cos(weidu(i ,j) *pi/ 180) ;d2 = cos(jingdu(i ,j) *pi/ 180) ;d5 = sin(jingdu(i ,j) *pi/ 180) ;d7 = sin(weidu(i ,j)
*pi/ 180) ;
    for k = 1 :91
        d3 = cos(wp (k) *pi/ 180) ;d4 = cos(jp (k) *pi/ 180) ;d6 = sin(jp (k) *pi/ 180) ;d8 = sin(wp (k) *pi/
180) ;
            d = ((d1 *d2 - d3 *d4)^2 + (d1 *d5 - d3 *d6)^2 + (d7 - d8)^2) ;
      if d > 0
        d = 1/ d ;dd = dd + d ;f = p(k) *d ;ff = ff + f ;          end          end
    P(i ,j) = ff/ dd; dd = 0 ;ff = 0 ;
    tt1 = P(i ,j) - A1(i ,j) ;tt2 = P(i ,j) - A2(i ,j) ;   sigma1 = sigma1 + abs(tt1) ;   sigma2 = sigma2 + abs(tt2) ;
      if P(i ,j) > 60.1      a = 6 ;      elseif P(i ,j) > 25.1      a = 5 ;      elseif P(i ,j) > 12.1      a = 4 ;
      elseif P(i ,j) > 6.1      a = 3 ;      elseif P(i ,j) > 2.6      a = 2 ;      elseif P(i ,j) > 0.1      a = 1 ;
      else    a = 0 ;      end
    if A1(i ,j) > 60.1      b = 6 ;      elseif A1(i ,j) > 25.1      b = 5 ;      elseif A1(i ,j) > 12.1      b = 4 ;
      elseif A1(i ,j) > 6.1      b = 3 ;      elseif A1(i ,j) > 2.6      b = 2 ;      elseif A1(i ,j) > 0.1      b = 1 ;
      else      b = 0 ;      end
    if A2(i ,j) > 60.1      c = 6 ;      elseif A2(i ,j) > 25.1      c = 5 ;      elseif A2(i ,j) > 12.1      c = 4 ;
      elseif A2(i ,j) > 6.1      c = 3 ;      elseif A2(i ,j) > 2.6      c = 2 ;
      elseif A2(i ,j) > 0.1      c = 1 ;      else      c = 0 ;      end
      switch abs(a - b)    case 1    ab = ab + 0.15276 *abs(a - b) ;    case 2    ab = ab + 0.47424 *abs(a - b) ;
        case 3      ab = ab + 0.68506 *abs(a - b) ;    case 4    ab = ab + 0.8 *abs(a - b) ;
        case 5      ab = ab + 0.9101 *abs(a - b) ;      case 6    ab = ab + abs(a - b) ;
        otherwise    ab = ab + abs(a - b) ;      end
      switch abs(a - c)    case 1    ac = ac + 0.15276 *abs(a - c) ;    case 2    ac = ac + 0.47424 *abs(a - c) ;
        case 3      ac = ac + 0.68506 *abs(a - c) ;      case 4      ac = ac + 0.8 *abs(a - c) ;
        case 5      ac = ac + 0.9101 *abs(a - c) ;      case 6      ac = ac + abs(a - c) ;
        otherwise    ac = ac + abs(a - c) ;      end
  end    end      end      sigma1 ,sigma2 ,ab ,ac
end
```

经过计算得到第一种预报方法的 w 值为 71.7 ,第二种预报方法的 w 值为 73.4 ,故认为第一种预报方法优于第二种预报方法 ,在考虑公众的感受后第一种预报方法的总误差为 3.66 ,第二种预报方法的总误差本为 3.70 ,计算结果也说明了预报方法一优于预报方法二。

2 参赛队解题中的典型错误

2.1 建模不合理

用数学建模来解决实际问题依赖于多种因素 ,必备的数学知识 ,对实际问题的深刻了解 ,抓住主要因素 ,作出正确的数学抽象的能力 ,模型的求解还依赖于计算技术的使用 ,通常要以计算机和数学软件包的熟练使用为技术手段。参赛队员既要有良好的高等数学、线性代数、概率统计等数学知识 ,也要熟悉基本的数学模型 ,如初等模型、微分方程模型、差分方程模型、线性规划模型、非线性规划模型、离散模型、概率统计模型等 ,还要掌握一些常用算法 ,计算机模拟蒙特卡罗算法 ,数据拟合算法、插值算法、规划类算法、图论算法、最优化理论的三大经典算法等等。由于在短时间内一个人要全面掌握以上知识是不现实的 ,当建模问题涉及到参赛队员知识、能力的薄弱之处时 ,建模不合理甚至建模错误也就在所难免。就本题而言 ,在考察预报误差与

建立评价准则函数时,某参赛队以整个区域上的两种预报方法所得面雨量(平均雨量)序列与观测站点区域上的面雨量(平均雨量)序列在正态分布的假设下进行假设检验,最后以两种预报方法的面雨量序列的方差大小为评判预报质量的优劣,该参赛队虽然独立思考了问题,但由于没有领会建立误差的插值思想,在数学知识不足时,导致建模竞赛失利;另外有一半参赛队被庞大数据难住,又没找到散乱数据的曲面拟合方法,只好对每个观测站点挑选距离最近的一个网格点,来比较误差,这种做法明显只取了 53 ×47 个网格点上的 91 个网格点,不合理之处也较明显。

2.2　建模方法不适当

数学建模竞赛题涉及多个领域中的实际问题,虽经适当简化,加工,但依然较难分门归类,建模也没有普遍适用的方法,需要根据具体问题选用合适的方法[6]。随着数学建模活动的不断发展、深入,建模竞赛题的难度也不断增大。一方面大规模、超大规模数据处理需要较好的硬件及数学软件,在参赛队所建模型相同时,模型求解就成了胜败关键(大型的混合 0 - 1 规划等模型求解就离不开正版软件);另一方面由于建模赛题完全来自实际问题,建模题所及内容的参考资料又很少,现有教材中的各种模型通常不能直接套用。同一个问题可以用不同的方法求解,不同的问题也可能用相同的方法建模,建模竞赛题通常没有标准答案,评判的标准是模型的合理性、论文表述的清晰程度、求解模型所用算法的优劣等等。由于 2005 数学建模 C 题的气象背景,尽管用实测值进行插值更易为人接受,但获奖论文普遍采用了气象部门的一贯做法,从等距网格点上找出距每个观测站点最近的四个网格点,用这些点上的预报值进行插值,得到观测站点的预报值,建模方法显得单一。另外文献的查找能力是建模竞赛的重要能力之一,但对于搜索到的参考资料,不能过于盲从,需要结合赛题考虑其方法的适用性。建立模型并求解后,要进行模型检验与评价,如发现模型不可用就要当机立断废弃,从头再来,否则得不到好的结果。本题中有的参赛队采用了某篇论文的计算公式,但在距离加权中,系数可能出现负数,该处被扣分,影响了成绩。有的参赛队使用了泰森多边形方法,但由于本题数据量过大,导致或者未能解完所建模型或者只挑选了部分数据进行计算,最后只得到末等奖。有的参赛队参照论文方法,用统计回归模型建模,但由于本题中经纬度跨度大,导致边缘区域误差太大,所建模型无法应用。

参考文献:

[1] 缪报通,陈发来. 径向基函数神经网络在散乱数据插值中的应用[J]. 中国科技大学学报,2001,31(2) P137
[2] 叶其孝. 大学生数学建模竞赛辅导教材(二)[M]. 长沙:湖南教育出版社,1997. 283.
[3] 殷浩,戴光明. 散乱数据可视化研究综述[J]. 微机发展,2005,15(7):8.
[4] 潘永地,徐为根. 沿海丘陵地区面雨量估算插值方法试验比较[J]. 气象科学 2005,25(2):124 - 132
[5] 张志涌. 精通 matlab6.5[M]. 北京:北京航空航天大学出版社,2003. 152 - 158
[6] 叶其孝. 大学生数学建模竞赛辅导教材(一)[M]. 长沙:湖南教育出版社,1993. 211.

Analysis on Problem C in CUMCM-2005

JIN Jian, ZHU Hui-jian
(Dept. of Math., Changshu Institute of Technology, Changshu 215500, China)

Abstract: This paper discusses the intention of the proposition of Problem C in CUMCM-2005: the Appraisal on Rainfall-Forecasting Methods. Based on the competition papers of our participating teams, the typical mistakes in solving the problem were analysed, and some essential points for the problem which is solved by curved surface fitting with the scattered data were suggested.

Key words: Shepard - interpolation; scattered data interpolation; curved surface fitting