气候资料缺测插补方法的对比研究*

张秀芝　孙安健

(国家气候中心,北京, 100081)

摘　要

采用均生函数正交筛选(MGF)和一维车贝雪夫多项式展开(CP)进行了年降水量各种缺测情况下资料的插补试验,并计算各种统计量,结果表明:无论何种缺测 MGF 法拟合或插补精度一般高于 CP 法,尤其对连续多年缺测和序列一开始便连续缺测更为明显;同一种方法 1a 缺测拟合精度高于多年缺测,连续 4—5a 缺测但在序列中处于不同的位置拟合结果差别不大,但多段同时缺测拟合精度低于一段缺测。

关键词:资料插补,均生函数,车贝雪夫多项式。

1　引　言

目前,气候变化和气候变异成为各国气候学家研究的热点问题。这就要求有一个长时期的均一的标准气候序列。但是,由于观测台站环境、观测仪器变更及战争等社会因素的影响,造成了长序列气候资料的非均一性及在某一时段出现间断。本文仅对器测时期以来的降水量缺测进行插补试验,旨在探索适应于中国历史资料插补的计算方法,希望能对历史资料标准化处理有所裨益。

2　方法及资料

设一维时间序列 $x(t)=\{x(1),x(2),\ldots,x(N)\}$,($N$ 为样本数),经变换

$$x'(t)=(x(t)-\bar{x})/\sigma \tag{1}$$

得到标准化序列 $x'(t)$。式中 $\bar{x},\sigma$ 分别为序列$\{x(t)\}$的平均值和均方差。对上述序列采用均生函数正交筛选和车贝雪夫多项式展开两种方法进行拟合计算。

2.1　均生函数正交筛选(Mean Generating Functoin MGF)

一维标准化序列 $x'(t)$ 的均生函数为

$$\bar{x}'_d(i)=(1/n_d)\sum_{l=0}^{n_d-1}x'(i+jd)\qquad i=1,2,\ldots,d\quad 1\leqslant d\leqslant M \tag{2}$$

式中 n_d 为满足 $n_d\leqslant[N/d]$ 的最大整数,$M=[N/2]$ 为不超过 $N/2$ 的最大整数。由于问题的焦点是资料中存在缺测,故需对上式中 n_d 的定义进行相应的修改,即在求间隔为 d 的 n 个数据的均值时,n 为 $x'(i+jd)$ 实有资料的个数,如当 $N=85,d=20,\bar{x}'_{20}(1)=$

* 初稿时间:1995 年 7 月 12 日;修改稿时间:1996 年 3 月 4 日。

$(1/n_{20})[x'(1)+///+x'(41)+x'(61)]$，由于$x'(21)$缺测，$\overline{x'}_{20}(1)$实际由3个数据求得，故$n_{20}=3$。当$d$接近$M$时，由于缺测的缘故，会导致$\overline{x'}_d$由少于2个的实测值求得，这时便停止$\overline{x'}_d$的计算。

重新定义后的式(2)仍具有其原有的特性：

(1)当序列的时间间隔d较小，即求均值的样本量较多时，缺少1－2个样本不会影响$\overline{x'}_d$的稳定性。

(2)当序列的时间间隔d较大，求均值的样本量较少，$\overline{x'}_d$的随机性变大，稳定性将有所下降。基于所研究的对象是近百年气候序列的缺测插补，只要d的取值不超过30，一般情况下$\overline{x'}_d$可由2个以上的数据相加平均而成，仍可保持其原有的特性。

均生函数$\overline{x'}(i)$的周期性延拓矩阵为

$$F=\begin{cases}\overline{x'} & \overline{x'}_2(1) & \overline{x'}_3(1) & \cdots & \overline{x'}_m(1)\\ \overline{x'} & \overline{x'}_2(2) & \overline{x'}_3(2) & \cdots & \overline{x'}_m(2)\\ & \cdots\cdots\cdots\cdots & & & \\ \overline{x'} & \overline{x'}_2(i2) & \overline{x'}_3(i3) & \cdots & \overline{x'}_m(im)\end{cases} \tag{3}$$

令f_2为正交化初始向量，对$f_3,f_4,\cdots,f_m$进行正交求得$M-1$个正交化序列$\tilde{f}_2,\tilde{f}_4,\cdots,\tilde{f}_m$，与$x(t)$建立线性模型

$$x(t)=\sum_{i=2}^{m}\tilde{\varphi}_i\tilde{f}_i(t)+e(t) \tag{4}$$

进入方程的均生函数个数则采用双评分准则确定[1]

原均生函数$f_i(t)$与$x(t)$的线性关系式为

$$\hat{x}(t)=\varphi_0+\sum_{i=1}^{k}\varphi_i f_i(t) \tag{5}$$

式中φ_i为原序列系数，k为进入方程参数个数。通过此式可得相应的缺测年份的计算值。

重新定义后的均生函数正交筛选方案，适用范围大大拓宽，既可用于气候序列的延伸预测，又可进行序列的缺测插补。

2.2 一维切贝雪夫多项式展开(Chebyshev Polynomials CP)

设在x轴上有T_0个等距点，在其上给定某气象要素的观测值$x'(t)$($t=1,2,\ldots,T_0$)，而在格点t_0处的要素值$x'(t_0)$缺测($t_0=1,2,\ldots,T_0$)。

将资料$x'(t)$在T_0个格点上用车贝雪夫多项式展开，即

$$\hat{x}'(t)=\sum_{k=0}^{k_0}\tilde{A}_k\tilde{\Phi}_k(t)\qquad(i=1,2,\ldots,T_0) \tag{6}$$

特别有

$$\hat{x}'(t_0)=\sum\tilde{A}_k\tilde{\Phi}_k(t_0) \tag{7}$$

其中

$$\tilde{A}_k=\sum_{\substack{t=1\\ t\neq t_0}}^{T_0}x'(t)\tilde{\Phi}_k(t)+x'(t_0)\tilde{\Phi}_k(t_0) \tag{8}$$

采用逐步逼近的方式求解上式，即

$$x'^{(v)}(t_0) = \sum_{k=0}^{k_0} \tilde{A}_k^{(v)} \tilde{\Phi}_k(t_0) \tag{9}$$

$$\tilde{A}_k^{(v)} = \sum_{\substack{t=1 \\ t \neq t_0}}^{T_0} x'(t)\tilde{\Phi}_k(t) + x'^{(v-1)}(t_0)\tilde{\Phi}_k(t_0) \tag{10}$$

迭代终值为

$$x'^{(\infty)}(t_0) = x'(t_0) + \hat{\varepsilon}/(1-\sigma) \tag{11}$$

其中

$$\hat{\varepsilon} = \hat{x}'(t_0) - x'(t_0)$$

$$\sigma = \sum_{k=0}^{k_0} \tilde{\Phi}_k^2(t_0)$$

插补公式是

当 $\hat{\varepsilon} > 0$ 时

$$x'_+(t_0) = x'^{(\infty)}(t_0) - (\bar{\hat{\varepsilon}}_+ /(1-\sigma)) \tag{12}$$

当 $\hat{\varepsilon} < 0$ 时

$$x'_-(t_0) = x'^{(\infty)}(t_0) - (\bar{\hat{\varepsilon}}_- /(1-\sigma)) \tag{13}$$

式中 $\bar{\hat{\varepsilon}}_+$ 和 $\bar{\hat{\varepsilon}}_-$ 分别为 $\varepsilon > 0$ 与 $\varepsilon < 0$ 时 ε 的平均值。

为了便于计算机操作，计算时假定若干个 T_0 值(如 3 ～ 15)，逐个代入分别计算其拟合误差，取误差最小所对应的 T_0 及 x'_∞ 作为插补值。由于计算是对序列依次在 T_0 个格点上展开，因此由观测起始年(如 1873 年) 至终止年 (如 1993 年) 按顺序计算或由终止年至起始年的逆序计算，其插补结果往往存在一定的差异，为了避免这种差异，本文除了序列一开始即连续缺测无法按顺序计算外，各种试验均采用双向((顺 + 逆)/2) 的计算结果。k_0 则根据文献[2] 的经验确定，ε 的符号取 1，插补值计算取 2。

利用上述两种方法，从年代较长资料较完整的上海、南京、昆明、天津、济南 5 站年降水量入手，计算、分析各有关统计量，比较两种方法的优劣。

3　计算结果分析

3.1　逐年拟合结果比较

表 1 为无缺测情况下 5 站各种拟合统计量，表中 σ 为序列均方差，$S = \sqrt{1/n\sum(\hat{x}(t) - x(t))^2}$ 为拟合均方误差。不难看出，全序列无缺测情况下拟合均方误差除了上海均生函数法(MGF)大于车贝雪夫法(CP)外，其余 4 站 MGF 均小于 CP，而且 S 值差别较大；拟合同号率表征计算值与实测值相对于序列均值的趋势符合情况，昆明、济南 MGF 高于CP，而南京、上海和天津则 CP 高于 MGF；为了更好地反应两种方法的拟合精度，尽可能消除由于南北方降水量差异较大拟合精度不宜比较的影响，进行拟合误差相对于序列标准差的计算 $rs = |(\hat{x}(t) - x(t))/\sigma|$；$rs$ 值小于等于 0.5(优) 出现频率，MGF 法 5 站均在 80% 以上，高于 CP，小于等于 1.0(良) 出现频率 MGF 法有 4 站达 100%，CP 法也有 3 站达 93% 以上，拟合相当不错；拟合误差相对于原序列 $re = |(\hat{x}(t)$

$-x(t))/x(t)|$（称之为相对误差）小于等于0.10（优）出现频率MGF法75%以上，CP法除天津外也达64%以上，小于等于0.20（良）的出现频率MGF约90%以上，CP达84%以上（天津除外）。图1为昆明两种方法拟合曲线，不难看出MGF计算值更接近于实测值，南京、天津、济南也表现出同样的规律，上海则CP拟合略好一些。

表1 无缺测逐年降水量拟合结果

		S	同号率	$rs\leqslant 0.5$	$rs\leqslant 1.0$	$re\leqslant .10$	$re\leqslant .20$	σ	n
上海	CP	111.5	87	76	93	78	95	209.6	
	MGF	157.8	84	80	90	79	90		121
南京	CP	117.5	93	76	95	70	91	232.1	
	MGF	71.1	90	91	100	88	99		89
昆明	CP	122.5	79	68	94	64	91	207.1	
	MGF	54.8	94	97	100	91	100		88
天津	CP	97.7	91	57	86	48	74	156.9	
	MGF	47.8	90	87	100	75	95		103
济南	CP	113.6	86	68	84	64	84	177.5	
	MGF	33.7	91	100	100	91	100		78

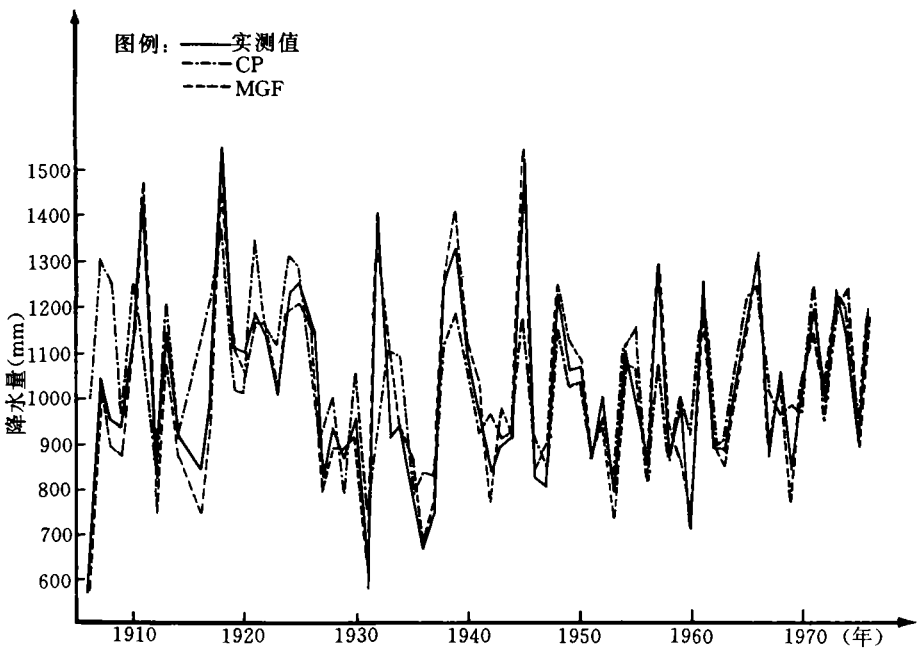

图1 昆明降水量拟合图

3.2 连续缺测多年拟合结果比较

由于战争的缘故，1930年代末至1940年代中国多数台站观测中断，缺测年数少则4—5a，多则十几年。为此，将5站中间年代连续10a资料去掉作假设缺测插补试验，各种统计结果列于表2。表中S为连续10a缺测全序列拟合误差，很明显MGF拟合效果好于CP，与表1无缺测情况下拟合误差相比，多年缺测S值MGF高出30mm左右，CP高出70—100mm。同号率为连续10 a假设缺测计算值与实测值符号相同的出现年数，

MGF 法 5 站同号年数一般 5－8a，CP 法 4－7a，除天津外其余 4 站均 MGF 多于 CP。拟合相对误差绝对值≤0.10 和≤0.20 出现年数 MGF 均多于 CP，拟合误差相对于序列标准差的 rs 值≤0.5 和≤1.0 出现年数也是如此。

表 2　连续 10a 缺测拟合结果

		S	同号率	$rs\leqslant 0.5$	$rs\leqslant 1.0$	$re\leqslant 0.10$	$re\leqslant 0.20$
上海	MGF	88.4	8	6	7	6	7
	CP	256.0	6	1	5	2	4
南京	MGF	85.5	6	4	9	6	9
	CP	177.5	4	6	7	6	6
昆明	MGF	88.5	7	4	8	4	8
	CP	235.1	4	1	7	2	6
天津	MGF	70.7	5	4	8	3	5
	CP	187.2	7	5	7	3	6
济南	MGF	65.7	6	6	7	5	6
	CP	145.7	5	4	7	2	5

3.3　缺测年数多少拟合结果比较

为了揭示缺测年数多少对计算精度的影响，分别同步计算了连续 4a 缺测、从连续 10a 缺测计算中取对应于连续 4a 缺测的计算值以及 1a 缺测的试验，各统计量列于表 3。很明显，无论拟合同号率还是 rs 及 re 值，拟合优的出现次数均 1a 缺测多于连续多年，但连续 4a 与连续 10a 缺测之间并无显著的差异。两种方法的计算精度仍以 MGF 为好。

表 3　各种情况缺测拟合结果比较

		同号率			$rs\leqslant 0.5$			$re\leqslant 0.10$		
		连续 4	10a 中 4	逐年	连续 4	10a 中 4	逐年	连续 4	10a 中 4	逐年
上海	MGF	3	4	3	2	2	1	2	2	1
	CP	3	2	3	2	1	3	2	1	3
南京	MGF	1	1	4	3	2	4	2	2	4
	CP	2	2	4	2	2	3	2	2	2
昆明	MGF	3	3	4	3	2	4	2	2	3
	CP	1	2	4	1	1	4	1	1	4
天津	MGF	2	3	4	4	3	4	2	3	4
	CP	1	3	4	3	3	4	1	3	3
济南	MGF	2	2	4	0	1	4	0	1	4
	CP	2	1	4	0	0	3	0	0	3

3.4　不同时段缺测拟合结果比较

造成观测资料残缺的原因很多，因而缺测在序列中的分布状况也较复杂，为了比较不同缺测分布对拟合效果的影响程度，分 3 段，即序列一开始连续缺测 5a(头 5a)，大约

20a 左右连续缺 4a(前 4a),序列中间连续缺 4a(中 4a),以及 3 种缺测同时出现(全缺),分别进行拟合计算.下面从两个方面进行分析讨论:

(1)对总体拟合结果的影响。表 4 和表 5 分别为各种缺测情况下 MGF 和 CP 法的拟合统计量,可以看出,处于不同时段单独 4a 或 5a 缺测情况下各方法自身的计算结果差别不大,即缺测处于不同时段对总体拟合结果影响不大,但两种方法相比则 MGF 显示出明显的优势,拟合均方误差 MGF 法一般仅有 CP 法的二分之一,拟合同号出现频率、$re \leqslant 0.10$ 以及 $rs \leqslant 0.5$ 出现频率 MGF 一般高出 CP 法 10%以上。当 3 个时段同时缺测,两种方法拟合情况均较一段缺测差,尤其是 CP 法,同号率、$re \leqslant 0.10$ 及 $rs \leqslant 0.5$ 出现频率减少 10%左右,与 MGF 相比频率差别更大,一般低于 MGF 法 20%--30%左右。3 个时段同时缺测拟合效果由图 2 反映的更为清楚,不缺测年份即使极端值都拟合的较好,尤其 MGF 法更接近于实测,但对于缺测的年份,特别是 1875,1931 年两种方法拟合的都不那么好。

表 4 不同位置缺测拟合均方误差

站名	头 5a		前 4a		中 4a		全缺	
	MGF	CP	MGF	CP	MGF	CP	MGF	CP
上海	76.4	115.2	59.4	114.9	71.6	143.7	88.6	144.2
南京	83.3	127.5	97.3	171.1	84.4	169.0	116.5	160.6
昆明	61.7	182.9	54.0	181.1	72.1	154.8	86.8	157.7
天津	40.4	98.3	59.2	119.0	48.7	114.8	71.3	133.7
济南	60.1	122.7	59.9	129.0	46.7	168.1	82.7	136.5

表 5 不同位置缺测拟合结果

		同号率				$rs \leqslant 0.5$				$re \leqslant 0.10$			
站名		头 5	前 4	中 4	全缺	头 5	前 4	中 4	全缺	头 5	前 4	中 4	全缺
上海	MGF	89	91	91	89	90	94	91	91	92	90	94	90
	CP	86	87	83	69	70	64	54	50		79	74	60
南京	MGF	85	89	85	87	90	87	90	83	85	88	88	83
	CP	91	76	76	56	56	67	60	48		65	55	44
昆明	MGF	97	93	91	89	90	95	94	84	95	89	95	84
	CP	61	63	77	56	37	31	53	40		43	33	41
天津	MGF	86	90	89	86	98	88	93	78	79	78	70	60
	CP	86	83	78	53	60	52	44	36		45	43	23
济南	MGF	90	91	92	86	94	92	97	83	90	85	68	64
	CP	82	75	77	60	61	61	41	39		57	48	31

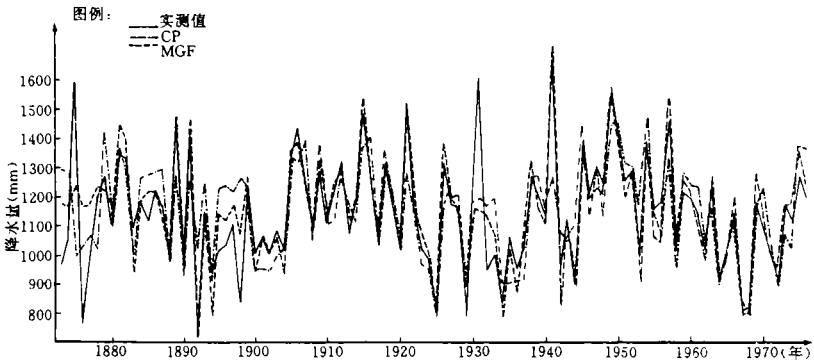

图 2　上海 3 段同时缺测拟合图

(2)缺测年份拟合结果比较。综合分析缺测年份各种拟合统计量发现，拟合优即 $rs \leqslant 0.5$ 出现年数 3 段分别缺测合计 MGF 法出现 5－6a，CP 法 4－6a，3 段同时缺测 MGF 法 4－6a，CP 法 2－6a；$re \leqslant 0.10$，3 段分别缺测合计出现年数 MGF 为 3－6a，CP 为 3－6a，3 段同时缺测 MGF 法 3－8a，CP 法一般 2－5a。拟合良好，即 $rs \leqslant 1.0$，3 种缺测单独出现时，13a 中合计 MGF 出现 8－10a，CP 5－8a，3 段同时缺测 MGF 7－10a，CP 4－9a；$re \leqslant 0.20$，3 段单独缺测 MGF 出现 7－10a，CP 6－9a，3 段同时缺测 MGF 8－10a，CP 5－9a(详见表 6)。从两种方法各自的拟合结果看，南京、济南 3 段同时缺测情况下的拟合结果还好于 3 段单独缺测，其余 3 站一般 1 段缺测好于全缺。拟合很差，即 $rs \geqslant 2.0$，13a 中 CP 法在上海、南京、昆明 3 站的拟合过程中分别各出现过 1a，且大多在开头 5a 缺测中，相比之下，MGF 法则很少出现，3 段同时缺测与各段单独缺测拟合精度差别不明显。从开头 5a 拟合结果看，拟合好和比较好的年数 MGF 均多于 CP。

表 6　缺测年份各站拟合结果

		$rs \leqslant 1.0$						$re \leqslant 0.20$					
站名		头 5	前 4	中 4	头 5	前 4	中 4	头 5	前 4	中 4	头 5	前 4	中 4
上海	MGF	3	3	3	3	4	3	4	3	2	2	3	3
	CP	2	3	3	1	1	1	3	3	3	1	2	2
南京	MGF	3	1	3	3	2	3	3	1	3	3	2	4
	CP	3	2	3	3	3	3	3	3	3	3	3	3
昆明	MGF	4	2	4	4	2	3	4	4	2	4	3	3
	CP	2	1	2	1	2	2	2	2	2	1	3	3
天津	MGF	2	2	4	2	1	2	2	1	4	2	2	4
	CP	2	1	3	0	1	3	2	1	3	0	1	1
济南	MGF	4	4	1	4	4	1	4	3	1	4	4	1
	CP	4	2	2	4	2	2	4	2	0	4	2	0

注:表中两栏的右边三列为 3 段同时缺测计算结果。

4 小 结

综上所述，MGF法对降水量各种缺测情况下的插补计算精度一般好于CP法，尤其对连续多年缺测和序列一开始便连续缺测的情况下更显示出优越性。造成这种差异的原因是，MGF法在均值生成函数时是按照一定的间隔d取值，只要从序列中能依序取得超过2个以上的实测值，便可依(3)式求得$x^{(t)}$值。而CP法是在T_0个格点上展开，且T_0个要素值中一般只有一个缺测值为好，从计算的5站看，T_0多数情况下为7－15之间，这对于序列一开始便连续多年缺测，顺续是无法进行计算的，只能采用逆序单向计算结果，势必产生一定的误差。

同一种方法，1a缺测拟合精度高于多年缺测；连续4－5a缺测与连续10a缺测拟合精度无明显差异；序列中缺测处于不同的位置拟合结果差别不大，但多段同时缺测拟合精度低于一段缺测。

对于特大涝年，两种方法计算结果都不太理想，有待进一步探讨。

参考文献

[1] 魏凤英等. 长期预报的数学模型及其应用. 北京：气象出版社，1990. 9－14，134－145.

[2] 周家斌. 车贝雪夫多项式及其在气象中的应用. 北京：气象出版社，1990. 146－148.

STUDY OF INTERPOLATION METHODS OF MISSING ANNUAL PRECIPITATION DATA FOR LONG-TIME SERIES

Zhang Xiuzhi Sun Anjian

(*National Climate Center, China Meteorological Administration, Beijing*, 100081)

Abstract

The interpolation experiments and statistics calculations of missing annual precipitation have been done using mean Generating Function (MGF) and one-dimensional Chebyshev polynomials (CP). The results have shown that MGF for fitting and interpolationaccuracy of missing annual precipitation are usually higher than CP, especially missing data with continuous many years and at beginning of series. In the case of using same method, accuracy of fitting for onlyone year missing is higher than that of many years. But there are not big difference between fitting results while 4—5a missing data in different location of data series. However accuracy of fitting forseveral parts missing at the same time is lower than that of only one part missing.

Key words: Data interpolation, Mean generating function, Chebyshev polynomials.