余予,李俊,任芝花,等. 标准序列法在日平均气温缺测数据插补中的应用[J]. 气象,2012,38(9):1135-1139.

标准序列法在日平均气温缺测数据插补中的应用*

余 予[1] 李 俊[2] 任芝花[1] 张志富[1]

1 国家气象信息中心,北京 100081

2 中国气象局预报与网络司,北京 100081

提 要: 利用标准序列法,对 1971—2000 年我国 2000 多个国家级地面气象站日平均气温进行了插补试验,并用交叉检验方法进行验证,对比了相关性最好和距离最近两种邻近站选取方案的插补结果。试验表明,相关性最好方案的插补精度优于距离最近方案,利用前一方案进行插补时,只需要选择与待插补站日平均气温序列相关性最高的 4 个邻近站参与计算即可。插补试验结果表明,平均绝对误差约为 0.42℃。插补值与实际观测值之间的绝对误差、均方根误差、两者之差在±0.5℃范围内的样本比例,均与邻近站平均距离呈较好的指数关系。

关键词: 标准序列法,日平均气温,缺测数据插补

Application of Standardized Method in Estimating Missing Daily Mean Air Temperature

YU Yu[1] LI Jun[2] REN Zhihua[1] ZHANG Zhifu[1]

1 National Meteorological Information Centre, Beijing 100081

2 Department of Forecasting and Information System, CMA, Beijing 100081

Abstract: Based on the daily mean air temperature from 1971 to 2000 observed by more than 2000 national surface stations in China, a standardized method was employed to carry out missing data estimation experiment, and the results were verified by cross-validation. Two schemes, the relation optimal scheme and the closest station scheme, which were both used to pick up the adjacent stations, were compared. It showed that the relation optimal scheme was better than the other, and only 4 adjacent stations that are most closely related to the estimated station were necessary for estimation. The results indicated that estimate values in average deviate from true values by 0.42℃. The absolute mean error and root mean square error between the estimation and the actural measurements and the sample ratio with the differences falling in ±0.5℃ were all shown good exponential relationships with the average adjacent station distance.

Key words: standardized method, daily mean air temperature, missing data estimation

引 言

长期完整的气温日值序列,是大气环流模式、陆面过程模型等模式所需要的输入参数,并且是进行气候统计分析和气候变化研究的基础。但是,由于站址迁移、台站撤并、观测仪器故障以及其他历史原因等,造成气温观测数据缺测或长时间序列中断,从而引起台站日气温序列不完整,而序列的不完整将对气候变化及其趋势研究、气候评估及影响评价产生影响[1]。因此,有必要对缺测的台站日气温资料进行估值计算,即对不完整的气温序列进行插补。

* 国家重点基础研究发展计划(2010CB951600)、中国科学院战略性先导科技专项子课题(XDA05090100)和国家气象信息中心青年基金项目"中国地面历史气温序列插补"共同资助

2011 年 8 月 4 日收稿; 2011 年 11 月 14 日收修定稿

第一作者:余予,从事气象资料处理分析与评估方面的工作. Email:yuyu@cma.gov.cn

早在 20 世纪 50 年代，么枕生[2]即提出了气温观测序列的订正问题。屠其璞[3]对气温序列的延长和插补进行了相关分析与研究。近年来，我国研究人员利用一维车贝雪夫多项式展开、线性回归、逐步回归、偏最小二乘回归等方法对我国部分地区的气温月、年值资料进行了恢复性试验[4-7]。此外，江志红等[8-9]、张永领等[10]对区域的气温场资料进行了插补研究。对台站观测的气温序列进行插补时，如果某站出现非连续的日气温缺测，可利用缺测日前后的气温数据进行插值计算，但是如果出现连续数日的气温缺测，应用该方法将造成较大误差[11]。标准序列法[12]是一种利用周边台站观测值进行插补的方法，它假设对于在同一气候区域内的所有站点，某日气温与该日多年平均气温的距平都是相似的。DeGaetano 等[13]对该方法进行了改进，并基于美国东北部近 400 站的气温资料，对日最高、最低气温缺测值进行了插补。王海军等[14]利用 DeGaetano 等的方法，利用湖北省蔡甸站周边 7 个台站的资料，对该站非缺测的日平均、最高、最低气温进行了插补试验，取得了较好的插补效果。

本文利用 DeGaetano 等的标准序列法，对 1971—2000 年我国 2000 多个国家级地面气象台站的日平均气温进行插补试验，并用交叉检验方法[15]验证其结果。使用标准序列法插补时，关键在于插补邻近站的选取。本文分别采用了“相关性最好”和“距离最近”两种邻近站选取方案，对比了两种方案的插补效果，同时对邻近站数的选择进行了讨论，并分析了不同台站其插补误差的大小与其邻近站平均距离之间的关系。

1 资料与方法

1.1 资 料

本文利用了中国地面 2400 台站气候资料日值数据集(2.0 版)[①]中的日平均气温数据，该数据已经通过气候界限值、台站气候极值、内部一致性和时间一致性等质量控制方法的检验。从地面 2400 台站日平均气温缺测率统计分析来看(见图 1)，1970 年前缺测相对较多，平均每月约有 37 站左右资料的缺失，20 世纪 90 年代由于某些台站未将地面月报文件上报国家气象信息中心，造成了部分站整月缺测。为了较好地检验插补效果，选用了 1971—2000 年 30 年日平均气温序列进行插补试验。

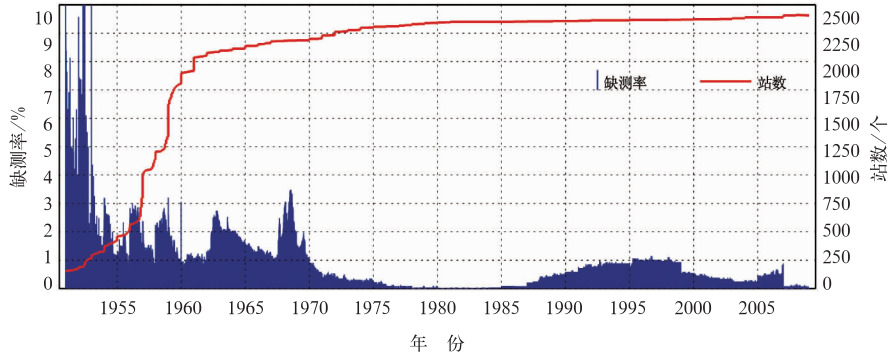

图 1 地面 2400 台站日平均气温缺测率和应有台站数变化

Fig. 1 The missing data rate of daily mean air temperature observed by 2400 national surface stations and its number change

1.2 插补方法

本文采用的插补方法为标准序列法[12]，计算时，首先假设待插补站某年中的第 i 日日平均气温缺测，然后利用邻近站日气温标准化距平，对插补站的气温值进行估计，该方法可表示为：

$$Z_j = \frac{X_j - \overline{X}_j}{S_j} \tag{1}$$

$$Z_{\text{avg}} = \frac{1}{n}\sum_{j=1}^{n} Z_j \tag{2}$$

$$X_i = Z_{\text{avg}} S_i + \overline{X}_i \tag{3}$$

式(1)～(3)中，Z表示标准化序列，Z_{avg}为邻站平均

[①] 中国气象科学数据共享服务网，http://cdc.cma.gov.cn.

标准化序列,j 代表第 j 个邻近站,X_j 为 j 站第 i 日日平均气温,$\overline{X}_j$ 和 S_j 分别为 j 站第 i 日日平均气温的多年(本文中即为 30 年)平均值和标准差,n 表示邻近站站数,X_i 表示第 i 日待插补日气温,$\overline{X}_i$ 和 S_i 分别为待插补站第 i 日日气温多年的平均值和标准差。

在插补前,首先建立了待插补站的邻近站表 Cls_sta0。在选择邻近站时,以待插补站为中心,在距其 220 km 的范围内进行搜索,并且备选站的海拔高度应满足:

当 $h_0 < 2500$ m 时,$|h-h_0| \leqslant 200$ m;

当 $h_0 \geqslant 2500$ m 时,$|h-h_0| \leqslant 500$ m;

式中,h_0、h 分别为待插补站和备选站的海拔高度。选择与待插补站距离最近的 20 个站为该站的邻近站,若邻近站数不足 20 个,以实际数为准。

在插补试验中,基于邻近站表 Cls_sta0 设计了相关性最好(RO 方案)和距离最近(CS 方案)两种不同的邻近站选取方案。采用 RO 方案时,需首先计算各邻近站与待插补站日平均气温序列的相关系数,然后选取相关性最高的 n 个邻近站参与计算;而 CS 方案则选取与待插补站距离最近的 n 个邻近站参与计算。

1.3　检验方法

本文采用交叉检验方法对上述两种邻近站选取方案的结果进行对比分析,并用平均绝对误差(MAE)、均方根误差($RMSE$)、插补值与实际观测值误差在±0.5℃范围内的样本比例(p)3 项指标来考查插补精度和插补效果。当邻近站数为 n 时,MAE_n、$RMSE_n$ 和 p_n 的计算公式分别为:

$$MAE_n = \frac{1}{m}\sum_{i=1}^{m} |x_{ei} - x_{oi}| \tag{4}$$

$$RMSE_n = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_{ei} - x_{oi})^2} \tag{5}$$

$$p_n = \frac{m_p}{m} \times 100\% \tag{6}$$

式中,x_{ei} 为第 i 日插补值,x_{oi} 为第 i 日实际观测值,m 为插补天数,m_p 为插补值与实际观测值误差在±0.5℃范围内的天数。MAE 和 $RMSE$ 的值越小,且比例 p 越大,则表明插补精度越高。

2　结果分析

针对 1971—2000 年 2088 个国家级台站的日平均气温,分别采用 RO 和 CS 两种邻近站选取方案进行了插补试验。基于插补结果,分别统计了每个待插补站其邻近站数为 n 时的 MAE、$RMSE$ 和比例 p 三项指标。

参考《中国气候总论》[16]将全国大致划分为 5 个气候区(图 2),在 5 个气候区中,分别随机选择了邻近站数相对较多的 1 个待插补站点,以上述统计指标 p 为例,给出了采用两种方案的不同插补精度,如图 3 和 4 所示。当采用 RO 方案时,5 个站的

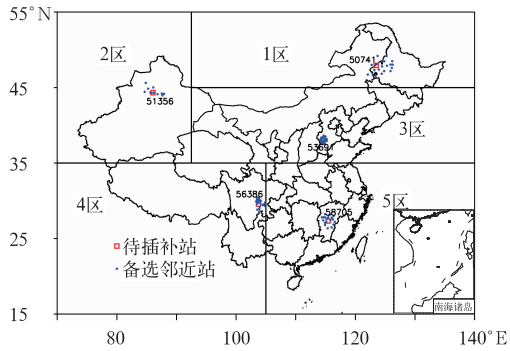

图 2　我国 5 个气候区划分和部分待插补站点及其邻近站分布

Fig. 2　Five climate regions in China as well as distributions of some estimated stations and their adjacent stations

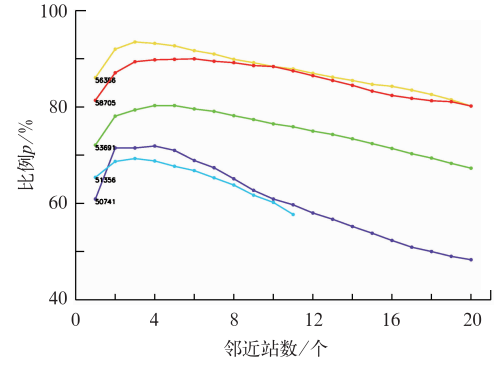

图 3　采用 RO 方案时比例 p 随邻近站数的变化

Fig. 3　Variations of proportion p with adjacent station number in relation to optimal scheme (RO)

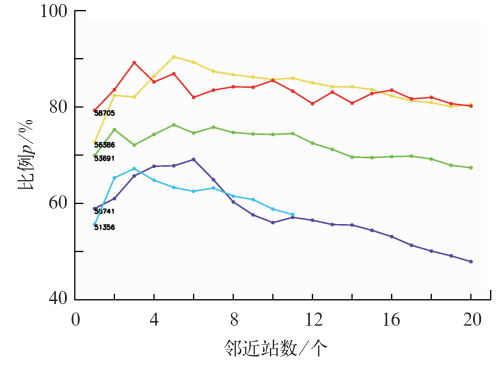

图 4　采用 CS 方案时比例 p 随邻近站数的变化

Fig. 4　Variations of proportion p with adjacent station number in relation to closest station scheme (CS)

比例 p 随邻近站数增加均先增大后减小，呈现单峰变化。采用 CS 方案时，比例 p 随邻近站数的变化不尽相同，53691 站的比例 p 随着邻近站数的增加变化不大，而 58705 站的比例 p 呈现出了多个峰值。对比发现，采用 CS 方案时 5 个站的比例 p 的最大值，比 RO 方案均有不同程度的减小。此外还可以看出，采用 RO 方案时，比例 p 在邻近站数为 3～4 时，即达到相对最大值，而采用 CS 方案时，比例 p 随邻近站数的变化没有规律性，达到相对最大值时的邻近站站数取值不固定。

对所有被插补的台站，分别统计了 *MAE*、*RMSE* 达到相对最小，比例 p 达到相对最大时的邻近站数取值，同样以比例 p 为例给出了两种方案的对比结果，如图 5 所示。当采用 RO 方案，约47.5%的台站当其邻近站数取为 3 或 4 时，比例 p 达到相对最大，约 88.1%的台站邻近站数取值在 2～6 之间比例 p 达到相对最大。而采用 CS 方案，约 72.3%的台站邻近站数取值在 2～6 之间时，比例 p 达到相对最大。由此可见，采用 RO 方案时的邻近站数取值相对比较集中。对比 *MAE* 和 *RMSE* 两项指标，也有类似的结论。

当邻近站数取为 4(2～6 的中值)时，对完成插补台站的 *MAE*、*RMSE* 和比例 p 进行了分区统计，并对比了两种方案的差异，结果在表 1 中给出。在 5 个不同气候区采用 RO 方案的 3 项指标的平均值均好于 CS 方案，并且各气候区中超过 85%的台站的 3 项指标 RO 方案优于 CS 方案。因此认为 RO 方案比 CS 方案的插补精度更高，在实际插补过程中应采用这一方案。对比不同气候区的插补精度可以看出，气候区 5 内的台站气温插补精度相对较高，*MAE* 为 0.340℃，而气候区 2 内的台站气温插补精度相对较低，*MAE* 为 0.672℃，这与不同气候区内台站的疏密有一定关系，将在第 3 节进行讨论。

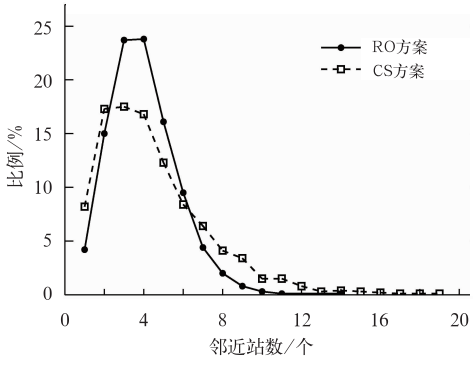

图 5 比例 p 为相对最大时的台站数占总数的百分比随邻近站数取值的变化

Fig. 5 Variations of station number ratio with adjacent station number when p obtains its relative maximum

表 1 两种方案的插补精度指标对比

Table 1 Comparison of 3 estimation indices in 2 schemes

指标	区域	均值		RO 方案优于 CS 方案的台站比例/%
		RO 方案	CS 方案	
MAE/℃	1 区	0.606	0.652	86.0
	2 区	0.672	0.735	91.9
	3 区	0.467	0.533	94.8
	4 区	0.519	0.569	94.2
	5 区	0.340	0.381	92.1
	全国	0.424	0.475	92.9
RMSE/℃	1 区	0.796	0.864	88.2
	2 区	0.880	0.966	91.9
	3 区	0.606	0.695	96.3
	4 区	0.676	0.744	95.1
	5 区	0.442	0.500	93.9
	全国	0.551	0.622	94.5
p/%	1 区	58.7	55.9	84.9
	2 区	54.0	51.0	88.7
	3 区	68.7	63.7	93.2
	4 区	65.2	61.7	94.2
	5 区	80.9	76.8	91.7
	全国	73.3	69.1	92.0

综上所述，在实际使用标准序列法进行气温插补时，应选用 RO 方案，并且对于备选邻近站数较多的待插补站点，并不需要利用所有的邻近站资料来进行插补，只需要选择与待插补站相关系数最高的 4 个邻近站资料参与计算，这样得到的插补结果具有相对较高的插补精度。

3 邻近站距离对插补效果的影响

从表 1 分区统计结果和图 3 可以看出，利用标准序列法对不同台站日平均气温的插补精度存在差异。一般来说，如果邻近站与待插补站的距离较近且海拔高度相差不大时，两者的日平均气温序列相关性较高，利用标准序列法做插补计算，可以得到较高的插补精度。假设插补计算某站第 i 日日平均气温时，最相关的 4 个邻近站与待插补站的平均距离为 d_i，则完成对 m 天日平均气温插补后的邻近站平均距离 d_{mean} 为：

$$d_{\mathrm{mean}} = \frac{1}{m}\sum_{i=1}^{m} d_i \tag{7}$$

对 2000 多个国家级台站应用 RO 方案进行插补试验后的 *MAE*、*RMSE* 和比例 p，这 3 项插补精度指标，与 d_{mean} 求取拟合关系式，如图 6 所示。可以看出，*MAE* 和 *RMSE* 随 d_{mean} 增加呈 e 指数增大，

而比例 p 随 d_{mean} 增加呈 e 指数减小,3 项插补精度指标与 d_{mean} 的相关性约为 0.8 左右,通过了显著性检验。这样,由图 6 中的 3 个拟合关系式,可以初步估计不同邻近站不同平均距离情况下的插补精度。但是,图 6 中的关系式是通过对大量插补样本进行统计后的结果拟合而成,不适用于对单个插值结果进行精度评估。

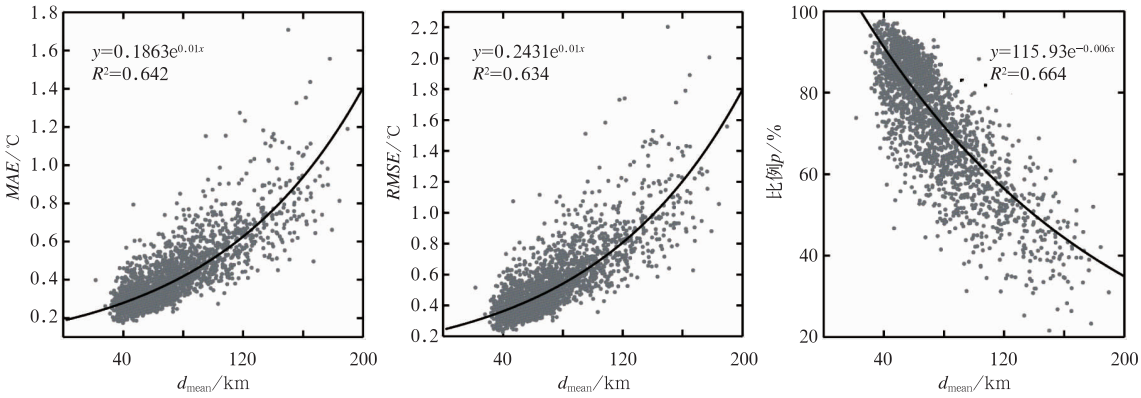

图 6　MAE、$RMSE$、比例 p 与 d_{mean} 之间的拟合关系

Fig. 6　Fitting relationships between MAE, $RMSE$ and p versus d_{mean}

4　结　论

基于我国地面 2000 多台站日值气候资料,利用标准序列法,进行了日平均气温插补试验,得到以下结论:

(1) 当利用标准序列法进行日平均气温插补时,采用相关性最好方案选取邻近站的插补结果优于距离最近方案。

(2) 对插补站某年第 i 日日平均气温进行插补时,选取与待插补站历史同期日平均气温序列相关性最高的 4 个邻近站参与插补计算,这样得到的插补值具有较高的精度。从我国地面 2000 多台站日平均气温插补试验结果来看,插补值与实际观测值的平均绝对误差为 0.424℃,均方根误差为0.551℃。

(3) 采用相关性最好方案时,MAE、$RMSE$、比例 p 与实际插补时所用的邻近站和待插补站的平均距离 d_{mean} 有较好的相关性,MAE 和 $RMSE$ 随 d_{mean} 的增加呈 e 指数增大,比例 p 随 d_{mean} 增加呈 e 指数减小。

参考文献

[1] Stooksbury D E, Idso C D, Hubbard K G. The effects of data gaps on the calculated monthly mean maximum and minimum temperatures in the continental United States: A spatial and temporal study[J]. J Climate, 1999, 12(5): 1524-1533.

[2] 么枕生. 中国境内农业指标温度的出现日期、持续日数与积算温度[J]. 地理学报, 1957, 23(2): 183-203.

[3] 屠其璞. 气温序列的延长和插补[J]. 气象, 1980, 6(5): 14-16.

[4] 张秀芝, 孙安健. 利用车贝雪夫多项式进行资料缺测插补的研究[J]. 应用气象学报, 1996, 7(3): 344-352.

[5] 涂诗玉, 陈正洪. 武汉和宜昌缺测气温资料的插补方法[J]. 湖北气象, 2001, 3: 11-13.

[6] 黄嘉佑, 刘小宁, 李庆祥. 夏季降水量与气温资料的恢复试验[J]. 应用气象学报, 2004, 15(2): 200-206.

[7] 李庆祥, 黄嘉佑, 鞠晓慧. 上海地区最高气温资料的恢复试验[J]. 热带气象学报, 2008, 24(4): 349-353.

[8] 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.

[9] 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.

[10] 张永领, 丁裕国, 高全洲, 等. 一种基于 SVD 的迭代方法及其用于气候资料场的插补试验[J]. 大气科学, 2006, 30(3): 526-532.

[11] Kemp W P, Burnell D G, Everson D O, et al. Estimating missing daily maximum and minimum temperatures[J]. J Climate Appl Meteor, 1983, 22(9): 1587-1593.

[12] Steurer P. Creation of a serially complete data base of high quality daily maximum and minimum temperature[M]. Washington D C: National Climate Center, NOAA, 1985, 21.

[13] DeGaetano A T, Eggleston K L, Knapp W W. A method to estimate daily maximum and minimum temperature observations[J]. J Appl Meteor, 1995, 34(2): 371-380.

[14] 王海军, 涂诗玉, 陈正洪. 日气温数据缺测的插补方法试验与误差分析[J]. 气象, 2008, 34(7): 83-91.

[15] Allen R J, DeGaetano A T. Estimating missing daily temperature extremes using an optimized regression approach[J]. Int J Climatol, 2001, 21(11): 1305-1319.

[16] 张家诚. 中国气候总论[M]. 北京: 气象出版社, 1991: 257-274.