基于动态时间规整的气温日值数据二次插补

周笑天[①②]，张茜茹[①②*]，郭庆燕[③]，陈益玲[①②]，周雪松[①②]，李长军[①②]，冯勇[①②]，张平[①②]

①山东省气象防灾减灾重点实验室，济南 250031；

②山东省气象数据中心，济南 250031；

③福建省气象信息中心，福州 350001

*联系人，E-mail: zhangqianru_m@163.com

XXXX-XX-XX 收稿，XXXX-XX-XX 接受

山东气象局重点科研项目(2021SDQXZ02)；山东省气象局青年科研基金项目(2021SDQN03)

摘要：气温作为研究气候演变最基础的物理量，其日值序列的完整性和准确性对于气候分析与评估工作有着重要的意义。近些年随着大量无人值守地面加密自动气象站的布设，不断出现随机站点和随机长度这种双随机特点的气象资料序列缺失，给气候分析和业务应用造成了不小的障碍。本文针对现有气象数据插补方案的不足，提出了一种全新的基于动态时间规整(Dynamic Time Warping, DTW)的气温日值数据二次插补方法。方法采用了一种实时的插补策略，主要技术内容包括：(1)利用一元线性回归方程将原始气温观测时间序列分解出拟合直线和残差曲线，并将二者重构组成新的气温序列；(2)给出了气温插补区的定义和插补条件；(3)提出了利用动态时间规整方法计算站点间距离的新模式。利用山东省 2021 年的气温实况数据对方法进行了双随机检验，检验结果表明：方法可以满足日平均气温、日最高气温和日最低气温数据的插补需求；在插补流程中采用 DTW 距离测度和二次插补的组合方法，插补效果优于目前常见的基于站点地理临近关系的组合方法；方法对于地形存在一定的敏感性，在平原或丘陵地区的插补效果优于山地地区。

关键词：气温日值；动态时间规整；重构；二次插补

气象资料序列的完整性是开展气候分析与评估的必要条件之一。随着我国精密气象监测系统建设的不断推进、地面气象观测站网布局逐步优化，区域观测盲区得到进一步消除，观测要素短板也实现了补足，为开展中小尺度灾害性天气监测预警业务和区域小气候特征分析创造了有利条件。与此同时，随着大量无人值守的地面加密自动气象站的布设，因仪器故障、通信中断、自然灾害等原因引起的观测中断或数据质量异常的发生概率大增，不断出现随机站点和随机长度这种双随机特点的气象资料序列缺失，给气候分析和业务应用造成了不小的障碍。

气温作为研究气候演变最基础的物理量，其日值序列的完整性和准确性对于气候统计与分析有着重要的意义，也是业内的关注焦点之一（Della-Marta P M et al.，2006；Hansen J et al.，2010）。目前，国内外对于气温日值缺失数据的插补，主要采用的方案是为数据缺失站点选择一个或者多个地理关系临近且气候相关性较高的数据参照站，并用参照站气温观测实值进行数据替代的方案。王海军等（2008）采用距离最短原则，以最小绝对偏差（Least Absolute Deviation, LAD）为目标函数求解模型参数，对处于平原地区的湖北蔡甸国家气象观测站日平均气温、日最高气温和日最低气温进行了插补试验和误差分析；此外，余予等（2012）采用标准序列法（Steurer P，1985）对1971-2000年我国2000余个国家级地面气象观测站日平均气温进行了插补试验，获得了较好的插补效果。以上插补方案虽然可以解决气温日值数据的插补，但是在具体实现时，需要周边参考站累年历史同期气温平均值和标准差（余予等，2012）以便计算气候相关性，这种限定条件对于

建站时间较短、迁建频繁、缺少长序列历史数据序列的无人值守气象站来说并不适用，时效上也较为滞后。

近些年，我国逐步建立并完善了包括气温要素在内的多源智能网格实况融合分析产品的小时级实时业务（李超等，2017；潘旸等，2018），客观上也促成了格点到站点间气温数据回插的实现（司鹏等，2022）。但是，由于气温实况格点是以ECMWF等数值预报产品为背景场，采用多重网格变分技术并融合地面站点的观测数据而生成（师春香等，2019；刘莹等，2021；张璐等，2017），格点数据产品的质量依然依赖于地面站点分布密度水平（龙柯吉等，2019；俞剑蔚等，2019）。当站点的观测数据在时间上和空间上出现随机性缺失时，关联时空范围内网格产品的稳定性也会受到极大影响（孙靖等，2021），从而极易造成数据回插异常。

因此，为了更科学和快捷地解决日益增多的双随机数据缺失问题，本文提出了一种全新的基于动态时间规整（Dynamic Time Warping，DTW）的气温日值缺失数据二次插补方法。方法采用了一种实时的插补策略，通过DTW距离计算、残拟分离和残拟重构等主要步骤，可以较为准确地实现观测站点气温日值数据的插补。

1　资料与方法

1.1 资料来源

本文中用于方法介绍、检验和分析的气温日值数据，来源于山东省气象大数据云平台“天擎”数据库中，并且全部经过数据质量控制（周笑天等，2012；王海军等，2014；叶小岭等，2019），时间范围涵盖2021年全年，包括日平均气温、日最高气温和日最低气温三种要素。

1.2 插补方法

1.2.1 气温时间序列的残拟分离和重构

气温日值缺失数据的插补，可以归结为气温时间序列中断点数据的预测与最优化求解问题，而一元线性回归作为天气气候业务中最基本也是最简单的预测分析方法之一（魏凤英，2007），可以将其原理和方法应用到插补过程当中。

设 $Y=[y_1,y_2,\cdots,y_N]$ 为气温观测值构成的时间序列，长度为 N，则可以建立一元线性回归方程，记作：

$$\hat{y}_i=a+bt_i,\quad i=1,2,\cdots,N。\tag{1}$$

式(1)中，t_i 为 y_i 所对应的时间，a 为回归常数，b 为回归系数，a 和 b 用最小二乘法估计。$\hat{Y}=[\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_N]$ 为拟合时间序列，其形态为一条直线，代表了气温的整体变化趋势。

观测值 y_i 到拟合值 $\hat{y}_i$ 之间的残差记为：

$$e_i=y_i-\hat{y}_i,\tag{2}$$

将(2)式变换形式可得

$$y_i=e_i+\hat{y}_i。\tag{3}$$

从(3)式中可以看出，观测值 y_i 可以由残差值 e_i 和拟合值 $\hat{y}_i$ 相加构成，即序列 $Y=[y_1,y_2,\cdots,y_N]$ 可以分解出残差序列 $E=[e_1,e_2,\cdots,e_N]$ 和拟合序列 $\hat{Y}=[\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_N]$，实现“残拟分离”；将残差序列 $E=[e_1,e_2,\cdots,e_N]$ 和拟合序列 $\hat{Y}=[\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_N]$ 的再次结合称为“残拟重构”。

1.2.2 气温插补区

设 $Y=[A,B,C]=[a_1,\cdots,a_N,b_{N+1},\cdots,b_{2N},c_{2N+1},\cdots,c_{3N}]$ 为气温观测时间序列，长度为 $3N$（N≥1），序列 $B=[b_{N+1},\cdots,b_{2N}]$（长度为 N）为序列 Y 中气温连续缺失部分，元素为空值。如果序列 B 的左邻序列 $A=[a_1,\cdots,a_N]$ 和右邻序列 $C=[c_{2N+1},\cdots,c_{3N}]$ 存在且元素完整，则称序列 B 是可插补的。插补后的序列 B 记为 $\hat{B}=[\hat{b}_{N+1},\cdots,\hat{b}_{2N}]$，插补后的序列 Y 记为：

$$\hat{Y}=[A,\hat{B},C]=[a_1,\cdots,a_N,\hat{b}_{N+1},\cdots,\hat{b}_{2N},c_{2N+1},\cdots,c_{3N}]。\qquad (4)$$

序列 A、序列 B 和序列 C 所在的时间区间分别称成为 A 区、B 区和 C 区，B 区也被称为插补区，序列 Y 和 $\hat{Y}$ 所在的时间区间被称作计算区间，如图 1 所示。

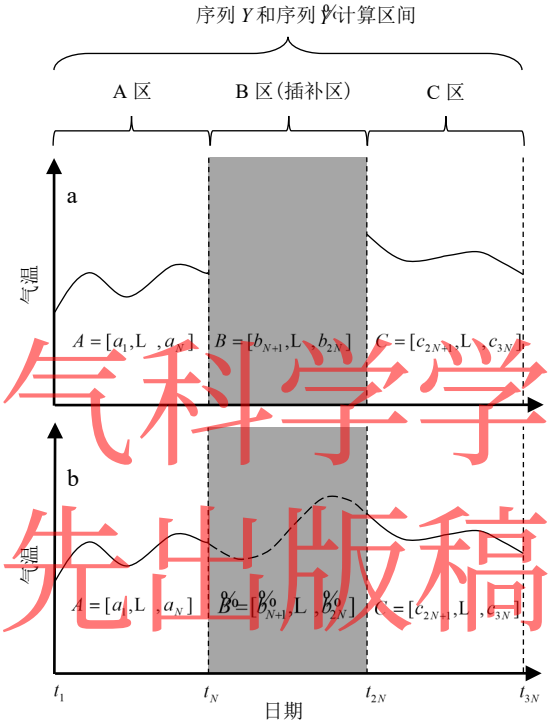

图1 气温插补区示意图：（a）插补前；（b）插补后

Fig.1 Schematic diagram of temperature interpolation zone: (a) before interpolation; (b)after interpolation

B区（插补区）及其左邻A区和右邻C区的设定，扩展了《地面气象观测规范》（中国气象局，2003）中缺测数据内插仅限于单时次的限定，从而可以有效解决“双随机”中的长度随机问题。

1.2.3 DTW距离与参照站

动态时间规整（Dynamic Time Warping, DTW）是时间规整和距离测度计算相结合的一种非线性归正计算方法，常被用在语音识别系统中（李正欣等，2014；闫宏宸等，2021）。该方法最大的特点就是可以通过路径规划来计算两个时间序列之间的最短累计距离，从而衡量二者的相似程度（Tormene T et al., 2008；周笑天等，2022）。

本文将动态时间规整算法原理应用于气温序列的相似度计算。设 $X=[x_1,x_2,\cdots,x_N]$ 和 $Y=[y_1,y_2,\cdots,y_N]$ 为两个不同站点的气温时间序列，长度为 N，如图 2 所示，上下两条实线分别代表序列 X 和序列 Y，动态时间

规整算法的路径规划过程即为从左向右刻画虚线的过程，该过程以(x_1,y_1)为起点，以(x_N,y_N)为终点，按照单调、连续的匹配原则，每向右行进一步画一条虚线，虚线两端分别匹配上下序列中的一个元素。

设行进至第 k 步时，虚线两端匹配的元素分别为x_i和y_j，则该步的步长距离为x_i和y_j气温之差的绝对值，记为$d(w_k)=\left|x_i-y_j\right|$。

设 L 为起点(x_1,y_1)至终点(x_N,y_N)刻画虚线的总数量，则$d(w_k)$最短累计步长，即为序列 X 和序列 Y 的 DTW 距离，记作：

$$R_{\mathrm{DTW}}(X,Y)=\min\{\sum_{k=1}^{L}d(w_k)\}。\tag{5}$$

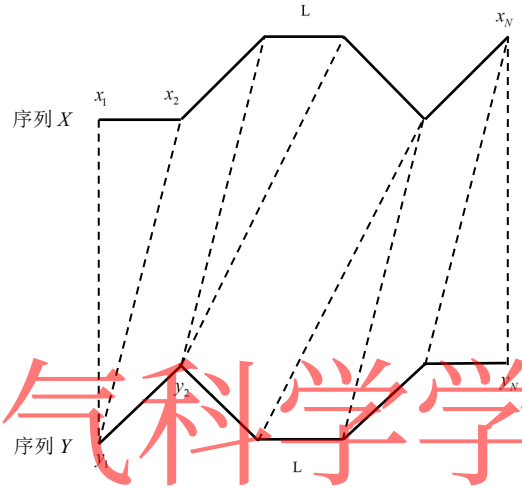

图2 序列X和序列Y的动态时间规整路径规划示意图

Fig. 2 Schematic diagram of DTW path planning for series X and series Y

结合气温插补区的设定，设站点 p 为需要进行数据插补的插补站，气温时间序列$Y^{(p)}=[A^{(p)},B^{(p)},C^{(p)}]$，$B^{(p)}$为可插补的数据缺失序列，如果存在站点 q ，其气温时间序列$Y^{(q)}=[A^{(q)},B^{(q)},C^{(q)}]$无数据缺失，且$Y^{(p)}$与$Y^{(q)}$的计算区间一致，则插补站 p 和站点 q 的 DTW 距离记作

$$D_{\mathrm{DTW}}(p,q)=R_{\mathrm{DTW}}(A^{(p)},A^{(q)})+R_{\mathrm{DTW}}(C^{(p)},C^{(q)})。\tag{6}$$

由式(6)可知，插补站 p 和站点 q 的 DTW 距离为两站 A、C 区序列的 DTW 距离之和。

将有限空间范围内的站点逐一与插补站 p 按照公式(6)计算 DTW 距离并排序，取 DTW 距离排序最小的站，称为参照站。

从以上定义可知，参照站是在 DTW 距离测度下，与插补站 p 气温序列最相似（DTW 距离最近)的站，参照站的气温序列适合作为插补数据源，执行后续的插补操作。

DTW 距离随着气温序列的变化而变化，因此是一种实时的、动态的距离计算方法。与传统的基于地理临近关系（如水平距离最近或者海拔高度最近）的参照站遴选方法不同，DTW 距离不受站网分布密度的影响，可以使遴选过程更加灵活和精准，更适用于解决“双随机”中站点随机的问题。

1.2.4 插补流程

综合上述的概念和方法，设站点 p 为需要进行数据插补的站，$B^{(p)}$ 为插补站 p 中连续数据缺失序列，则对 $B^{(p)}$ 的插补过程描述如下（流程如图 3 所示）：

第一步，输入插补站 p 的 $B^{(p)}$ 序列，内容为空值，长度为 N；

第二步，确定 $B^{(p)}$ 序列的左邻序列 $A^{(p)}$ 和右邻序列 $C^{(p)}$，长度都为 N，合并得到序列 $Y^{(p)}=[A^{(p)},B^{(p)},C^{(p)}]$；

第三步，从插补站 p 周边临近站中，按照 DTW 测度得到参照站 q，参照站 q 的序列 $Y^{(q)}=[A^{(q)},B^{(q)},C^{(q)}]$；

第四步，将参照站 q 的 $B^{(q)}$ 直接嫁接于插补站 p 的 $A^{(p)}$ 和 $C^{(p)}$ 之间，用以替换 $B^{(p)}$，形成插补站 p 第一次插补序列 $\%^{(p)}=[A^{(p)},\%^{(p)},C^{(p)}]$，$\%^{(p)}=B^{(q)}$，完成一次插补；

第五步，分别对参照站 q 的 $Y^{(q)}=[A^{(q)},B^{(q)},C^{(q)}]$ 和插补站 p 的一次插补序列 $\%^{(p)}=[A^{(p)},\%^{(p)},C^{(p)}]$ 建立时间为自变量的一元线性回归方程，$Y^{(q)}$ 对应拟合序列 $\hat{Y}^{(q)}=[\hat{A}^{(q)},\hat{B}^{(q)},\hat{C}^{(q)}]$ 和残差序列 $E^{(q)}=[E_A^{(q)},E_B^{(q)},E_C^{(q)}]$，$\%^{(p)}$ 对应拟合序列 $\hat{\%}^{(p)}=[\hat{A}^{(p)},\hat{\%}^{(p)},\hat{C}^{(p)}]$ 和残差序列 $E\%^{(p)}=[E_A^{(p)},E\%_B^{(p)},E_C^{(p)}]$，实现残拟分离；

最后，将序列 $E_B^{(q)}$ 和序列 $\hat{\%}^{(p)}$ 相加，残拟重构得到序列 $B_r^{(p)}$，$B_r^{(p)}$ 即为插补站 p 数据缺失序列 $B^{(p)}$ 的二次插补结果。

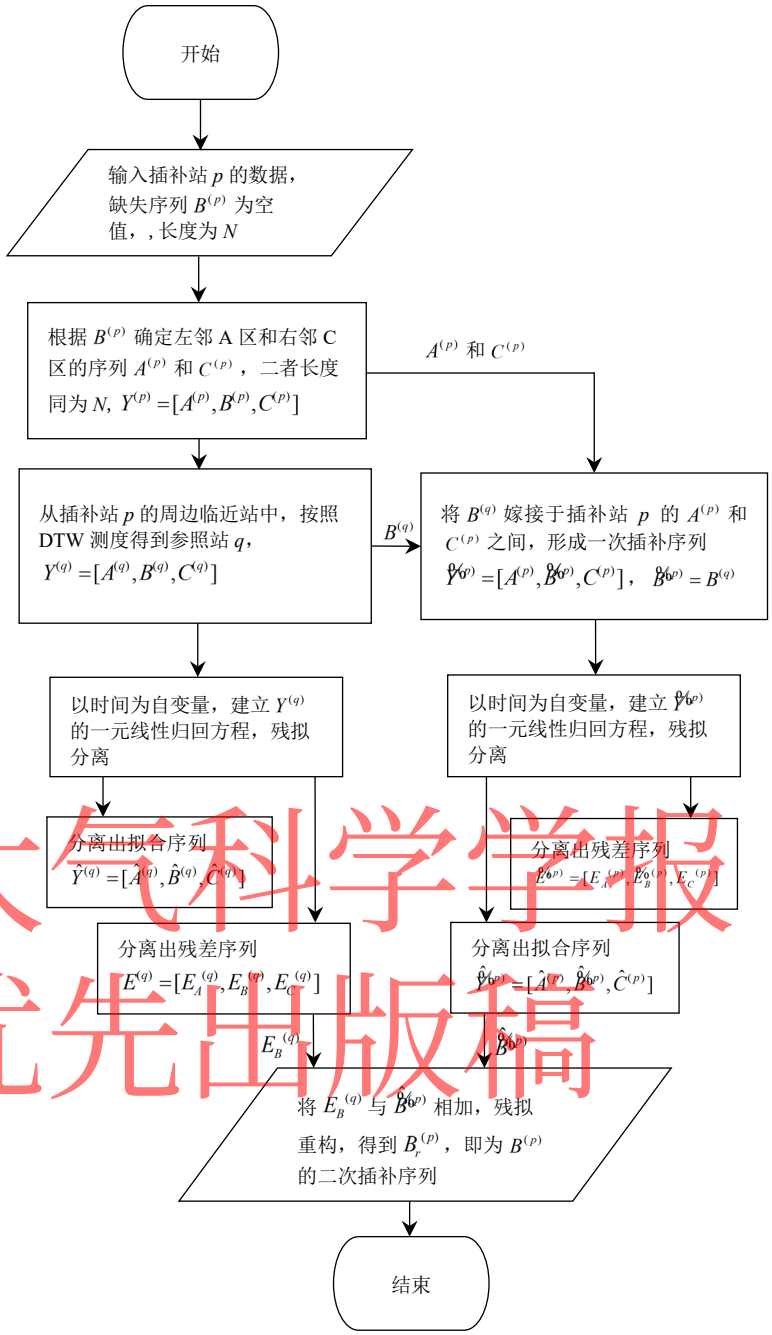

图3 基于动态时间规整的气温日值数据插补流程

Fig. 3 Flow chart of daily temperature interpolation based on DTW

从图 3 插补方法总体流程可以看出，方法采用了数据嫁接的操作方式，共分为两个插补阶段，一次插补为 $B^{(q)}$ 对 $B^{(p)}$ 的直接替换，二次插补是在一次插补的基础上，对两站的气温序列进行残拟分离和残拟重构后，得到的插补结果。

需要说明的是，插补流程中第三步是以 DTW 距离测度为基准选定参照站，而当以地理距离测度为基准（如水平距离最近或者海拔高度最近等）选定参照站时，则应以相应距离计算方法代替。

2 方法检验与分析

2.1 检验设计

为了验证方法对实况数据插补的有效性，同时也为了使得检验过程更加完备，本文对检验条件和检验内容设计如下：

(1)检验包括日平均气温、日最高气温和日最低气温三种要素在内的气温日值数据；

(2)根据山东省气象地理一级区划规则，分别在鲁西北（以平原地形为主）、鲁中（以山地地形为主）、鲁南（以平原丘陵地形为主）和半岛（以丘陵和山地海岸地形为主）四个地区，每个地区中随机选择10个无人值守的地面气象观测站点作为插补测试站（如图4所示），在兼顾地形特征的同时满足站点分布的随机性要求；

(3)每个插补测试站的插补区起止时间随机产生，以满足插补长度随机性要求；

(4)检验采用观测真值置空的方式模拟插补区的缺失数据，并在插补结束后计算插补值与观测真值之间的误差以评估插补效果。具体的误差评估指标包括均方根误差（RMSE）：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\%_i-y_i)^2}\ , \tag{7}$$

和平均绝对误差(MAE)：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\%_i-y_i\right|\ 。 \tag{8}$$

式(7)和式(8)中，$\%_i$为插补值，y_i为观测真值，n为插补总样本数（站次数），RMSE和MAE的数值越小，表示插补值和观测真值之间的总体差距越小，插补效果越好；

(5)在插补流程执行过程中，采用条件组合覆盖法，记录插补阶段（三次、二次插补）和距离测度（DTW距离、水平距离和海拔高度）的各选项组合所能产生的全部插补结果。

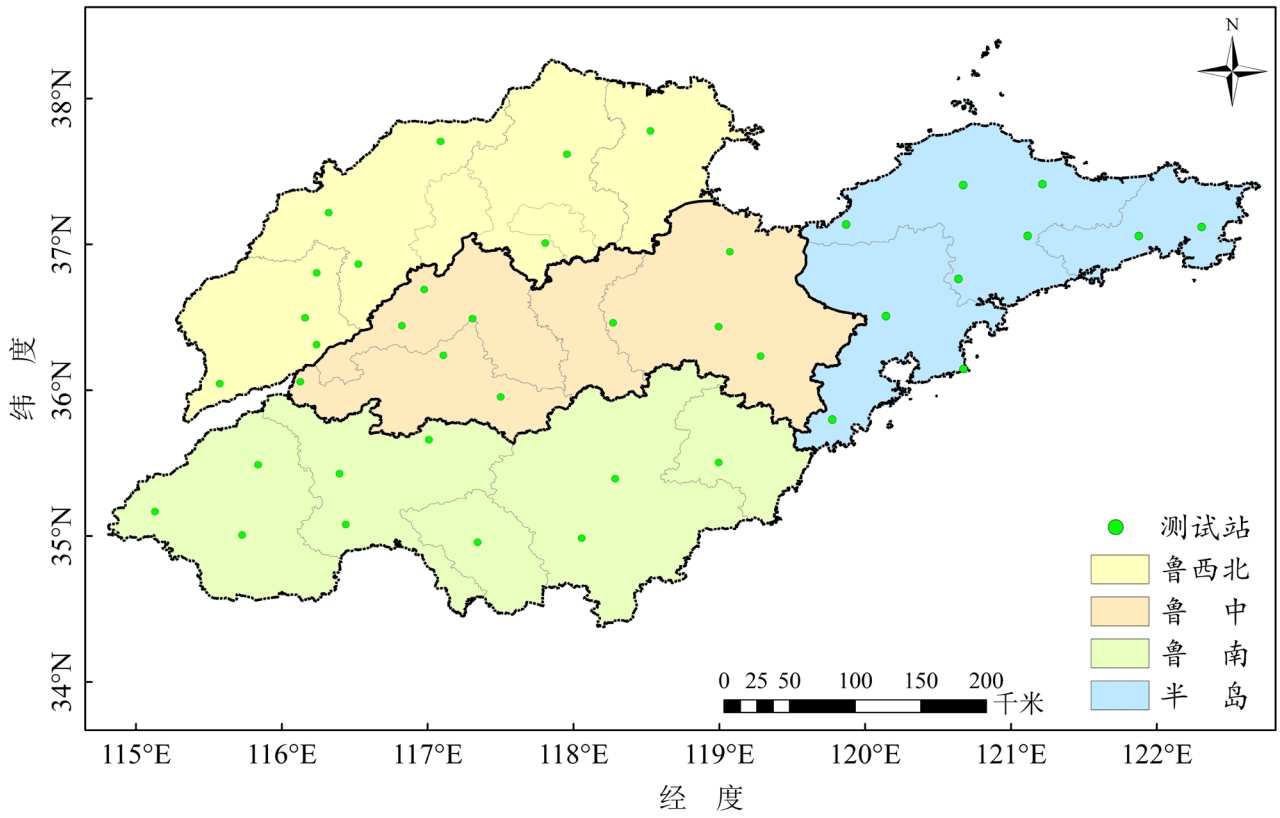

图4 山东省四个地区双随机测试站点分布图

Fig.4 Distribution of double random testing stations in four regions of Shandong province

2.2 插补实例

本文挑选了具有代表性的插补实例进行插补过程检验。选择的插补站 D0122（36.1458°N,120.6744°E）站址位于山东省青岛市境内，海拔 174 米，东、南两面临海，插补站 D0122 及其临近无人值守气象站网分布如图 5 所示。下文以 D0122 站日平均气温在 DTW 距离测度下的插补为例，验证插补流程的可行性。

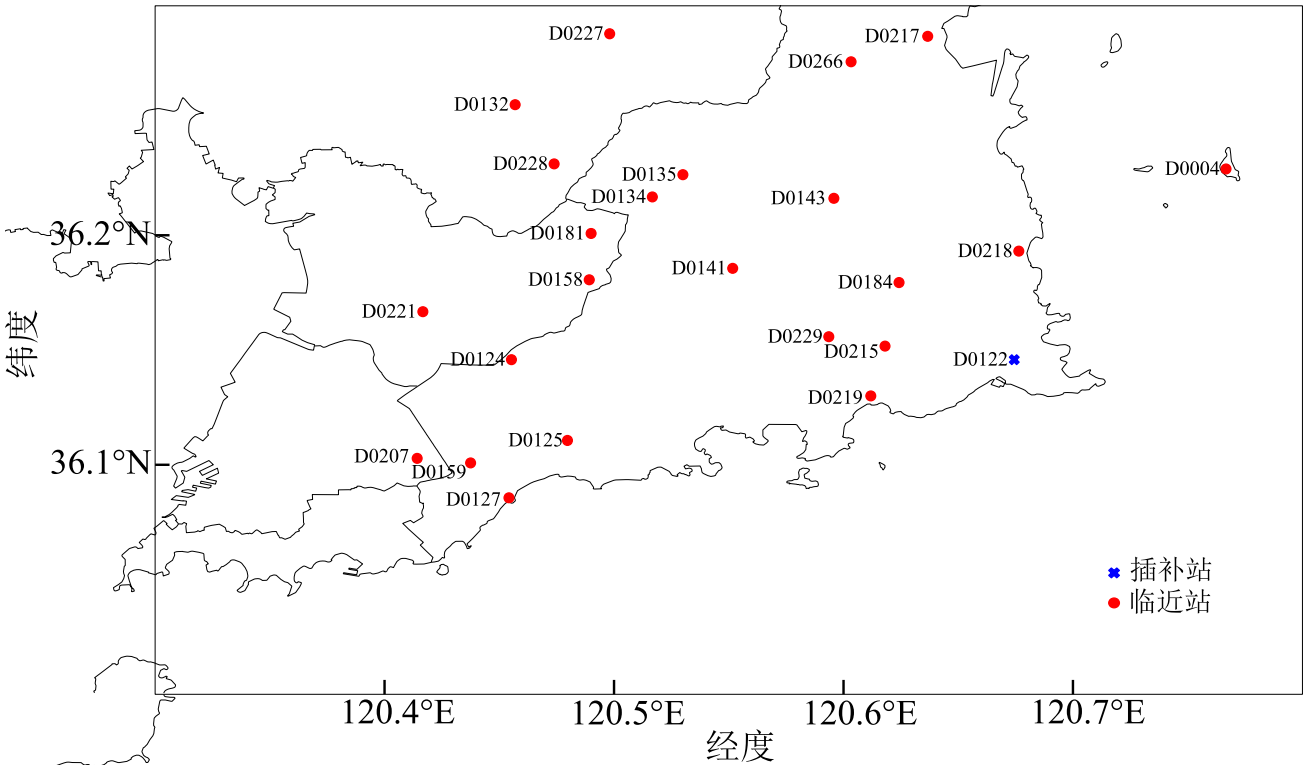

图5 插补站D0122及其临近无人值守气象站网分布图

Fig.5 Distribution of interpolation station D0122 and its adjacent unmanned meteorological station network

插补站 D0122 的日平均气温缺失区间为 2021 年 7 月 1 日至 9 月 30 日，长度为 92 天，该时间段即为需要进行数据插补的 B 区（插补区）。从 B 区的区间范围可以确定 B 区的左邻 A 区的区间范围为 2021 年 3 月 31 日至 6 月 30 日，B 区右邻 C 区的区间范围为 2021 年 10 月 1 日至 12 月 31 日，A 区和 C 区内的日平均气温数据完整，区间长度同为 92 天（分区如图 6 所示）。

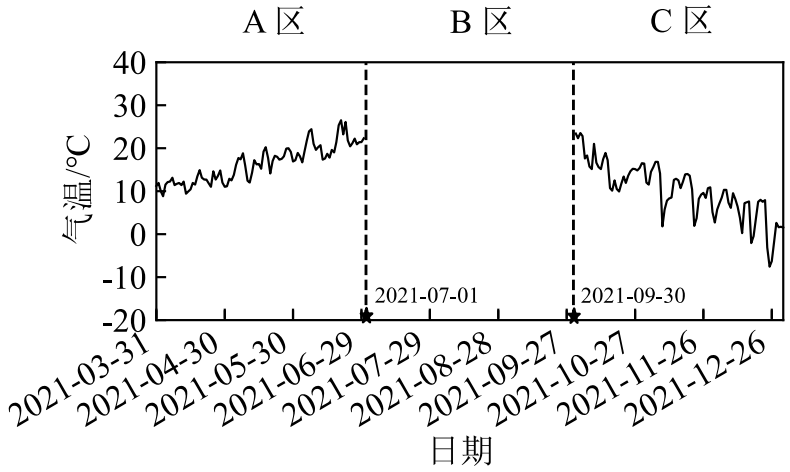

图 6 插补站 D0122 日平均气温序列分区（A 区日期区间为 2021 年 3 月 31 日至 6 月 30 日，B 区日期区间为 2021 年 7 月 1 日至 9 月 30 日，C 区日期区间为 2021 年 10 月 1 日至 12 月 31 日。虚线为区间分界线，五星标识为分界日期）

Fig. 6 Zone division of daily mean temperature series of interpolation station D0122 (The date range of zone A is from March 31 to June 30, 2021. The date range of zone B is from July 1 to September 30, 2021. The date range of zone C is from October 1 to December 31, 2021.The dotted lines are the zone boundaries, and the five-stars mark the date of demarcation)

在 DTW 距离测度下，遍历插补站 D0122 的临近站（图 5），按照公式（6）对临近站逐一计算日平均气温的 DTW 距离（如表 1 所示）。根据表 1 中的 DTW 距离排序，将 DTW 距离最小的站确定为参照站，站号 D0181，距离为 218.7℃。

表1 插补站D0122与临近站日平均气温的DTW距离

Table 1 DTW distance of daily mean temperature between interpolation station D0122 and its adjacent stations

临近站	D0215	D0218	D0184	D0219	D0229	D0143	D0141	D0004	D0135	D0266	D0217	D0134
DTW 距离/℃	331.3	287.2	517.4	326.9	276.0	318.5	291.7	224.5	322.7	293.3	351.5	366.0
临近站	D0158	D0181	D0125	D0124	D0228	D0127	D0159	D0227	D0132	D0221	D0207	
DTW 距离/℃	305.6	218.7	275.6	297.1	301.4	282.9	265.2	268.5	294.8	250.9	239.5	

在选定 D0181 为参照站后，我们结合图 7，对后续插补过程描述如下：

(1)将参照站 D0181 的 B 区（图 7a’中的 B 区），直接嫁接于插补站 D0122 的 B 区（图 7a 中的 B 区红色曲线段），此为一次插补；

(2)分别对一次插补序列（图 7a）和参照站 D0181 序列（图 7a’）计算一元线性归回方程，残拟分离，获得相应的拟合线和残差线（图 7b 和图 7b’）；

(3)将一次插补的拟合线（图 7b 中的 B 区取值部分）与参照站 D0181 残差线（图 7 b’的 B 区取值部分）相加重构，并再次嫁接于插补站 D0122 的 B 区（图 7c 中 B 区紫色曲线段），完成二次插补。

图 7c 中 B 区气温序列即为图 6 中 B 区缺失的日平均气温序列的最终插补结果，插补过程至此结束。

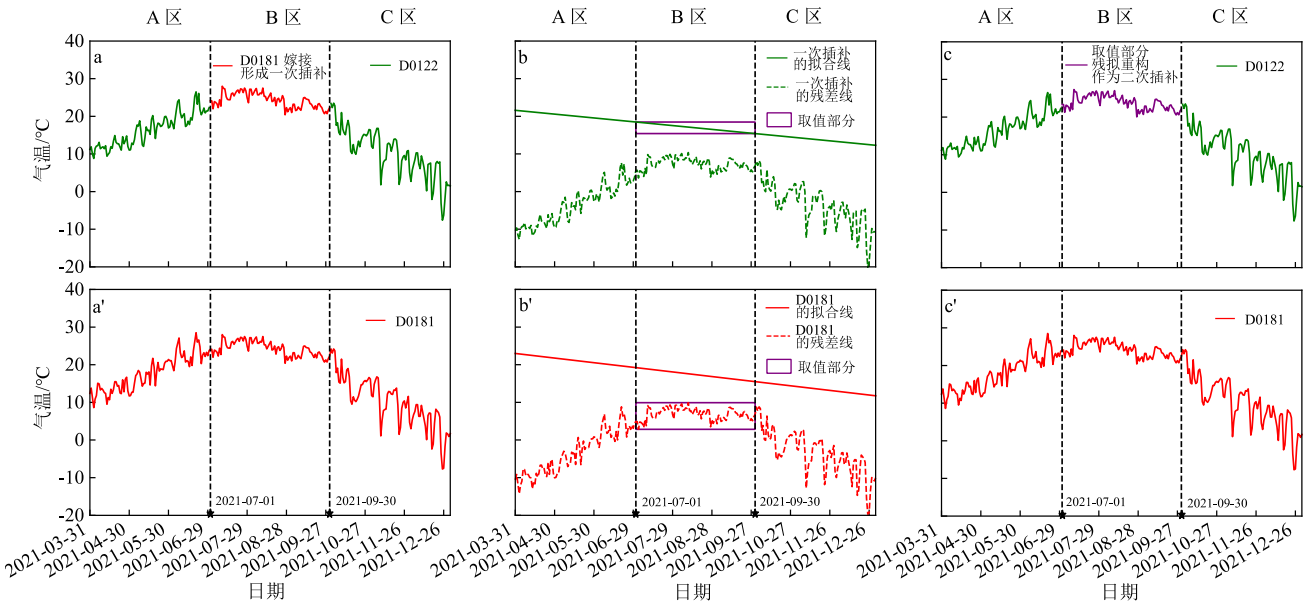

图 7 日平均气温的插补过程：(a)一次插补；(b)和(b')残拟分离；(c)二次插补；(a')和(c')参照站序列（D0122为插补站，D0181 为参照站。A 区日期区间为 2021 年 3 月 31 日至 6 月 30 日，B 区日期区间为 2021 年 7 月 1 日至 9 月 30 日，C 区日期区间为 2021 年 10 月 1 日至 12 月 31 日。黑色虚线为区间分界线，五星标识为分界日期）

Fig. 7 The interpolation process of daily mean temperature: (a) the primary interpolation; (b) and (b') the residual-fit separation; (c) the twice interpolation; (a') and (c') the series of reference station(D0122:the interpolation station; D0181:the reference station. The date range of zone A is from March 31 to June 30, 2021. The date range of zone B is from July 1 to September 30, 2021. The date range of zone C is from October 1 to December 31, 2021.The black dotted lines are the zone boundaries, and the five-stars mark the date of demarcation)

2.3 检验结果与分析

本文将日平均气温、日最高气温和日最低气温三种要素在山东省四个地区的双随机检验结果分别汇总在表2-表4中，并按地区分组对RMSE和MAE指标最小值进行了标注。

表2 日平均气温在山东省四个地区的双随机插补检验结果

Table 2 Double random interpolation and test results of daily mean temperature in four regions of Shandong province

距离测度	插补阶段	鲁西北地区		鲁中地区		鲁南地区		半岛地区	
		RMSE/°C	MAE/°C	RMSE/°C	MAE/°C	RMSE/°C	MAE/°C	RMSE/°C	MAE/°C
DTW距离	一次插补	0.446	0.354	0.828	0.704	0.406*	0.313	0.598	0.493
	二次插补	0.408*	0.348△	0.680*	0.554△	0.406*	0.312△	0.546*	0.438△
水平距离	一次插补	0.542	0.444	1.723	1.581	0.561	0.453	0.667	0.561
	二次插补	0.459	0.365	1.033	0.898	0.424	0.332	0.561	0.470
海拔高度	一次插补	0.591	0.480	0.973	0.828	0.686	0.569	0.686	0.544

	二次插补	0.528	0.422	0.757	0.606	0.480	0.378	0.634	0.507

注：*符号标注同一地区中RMSE指标最小值，△符号标注同一地区中MAE指标最小值。

表3 日最高气温在山东省四个地区的双随机插补检验结果

Table 3 Double random interpolation and test results of daily maximum temperature in four regions of Shandong province

距离测度	插补阶段	鲁西北地区		鲁中地区		鲁南地区		半岛地区	
		RMSE/℃	MAE/℃	RMSE/℃	MAE/℃	RMSE/℃	MAE/℃	RMSE/℃	MAE/℃
DTW距离	一次插补	0.738	0.571	1.489	1.284	0.880	0.642	1.019	0.764
	二次插补	0.725*	0.561△	1.029*	0.802△	0.859*	0.633△	0.960*	0.732△
水平距离	一次插补	0.907	0.704	2.277	2.061	0.976	0.803	1.161	0.912
	二次插补	0.852	0.658	1.301	1.076	0.859*	0.676	1.042	0.813
海拔高度	一次插补	0.939	0.753	1.751	1.468	1.102	0.784	1.533	1.217
	二次插补	0.843	0.669	1.225	0.953	0.940	0.724	1.306	1.051

注：*符号标注同一地区中RMSE指标最小值，△符号标注同一地区中MAE指标最小值。

表4 日最低气温在山东省四个地区的双随机插补检验结果

Table 4 Double random interpolation and test results of daily minimum temperature in four regions of Shandong province

距离测度	插补阶段	鲁西北地区		鲁中地区		鲁南地区		半岛地区	
		RMSE/℃	MAE/℃	RMSE/℃	MAE/℃	RMSE/℃	MAE/℃	RMSE/℃	MAE/℃
DTW距离	一次插补	0.509	0.382	0.950	0.745	0.538	0.377	0.722	0.547
	二次插补	0.487*	0.363△	0.900*	0.693△	0.515*	0.355△	0.712*	0.542△
水平距离	一次插补	0.688	0.539	1.749	1.531	0.693	0.519	1.070	0.850
	二次插补	0.585	0.456	1.305	1.108	0.591	0.444	0.909	0.739
海拔高度	一次插补	0.859	0.671	0.924	0.721	0.989	0.777	0.804	0.599
	二次插补	0.711	0.583	0.900*	0.698	0.729	0.551	0.778	0.627

注：*符号标注同一地区中RMSE指标最小值，△符号标注同一地区中MAE指标最小值。

根据指标排序情况可知：

(1)对于日平均气温（表2），在鲁西北、鲁中和半岛地区，DTW距离测度下的二次插补结果在RMSE指标和MAE指标上均表现最优；在鲁南地区，DTW距离测度下的二次插补结果虽然在RMSE指标上与一次插补表现持平（RMSE=0.406℃），但在MAE指标上表现更优（MAE=0.312℃）。

(2)对于日最高气温（表3），在鲁西北、鲁中和半岛地区，DTW距离测度下的二次插补结果在RMSE指标和MAE指标上均表现最优；在鲁南地区，DTW距离测度下的二次插补结果虽然在RMSE指标上与水平距离测度下的二次插补表现持平（RMSE=0.859℃），但在MAE指标上表现更优（MAE=0.633℃）。

(3)对于日最低气温（表4），在鲁西北、鲁南和半岛地区，DTW距离测度下的二次插补结果在RMSE指标和MAE指标上均表现最优；在鲁中地区，DTW距离测度下的二次插补结果虽然在RMSE指标上与海拔高度测度下的二次插补表现持平（RMSE=0.900℃），但在MAE指标上表现更优（MAE=0.693℃）。

综合检验结果，可以看出，日平均气温、日最低气温和日最高气温三种要素均能成功完成双随机条件下

的数据插补检验。与此同时，三种要素在插补流程中采用DTW距离测度和二次插补的组合方法，插补效果均优于基于水平距离、海拔高度的插补组合方法。

分别将表2-表4中四个地区的RMSE指标最小值作为比较对象，进一步研究分析不同地区的插补效果。从图8可以看出，日平均气温、日最高气温和日最低气温三种要素的插补误差，具备近似的分布特征，均在鲁中地区最高，半岛地区其次，鲁西北和鲁南地区最低。考虑到鲁中地区和半岛地区多以山地地形为主，而鲁西北和鲁南地区多以平原和丘陵地形为主，因此可以认为，复杂地形是干扰和降低本文所提方法插补效果的一个重要因素。

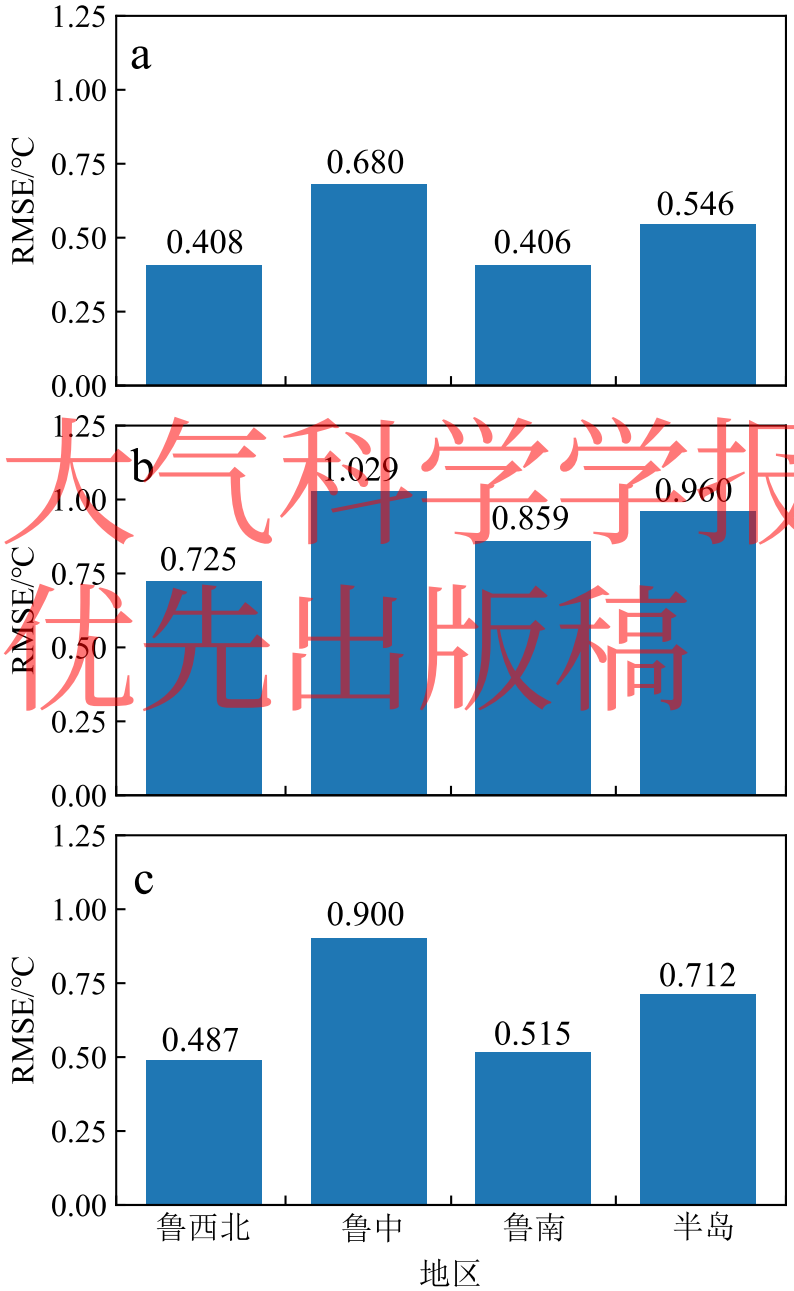

图8　山东省四个地区RMSE指标最小值：(a)日平均气温；(b)日最高气温；(c)日最低气温

Fig.8 Minimum RMSE values in four regions of Shandong province：(a) daily mean temperature; (b) daily

maximum temperature; (c) daily minimum temperature

3 结论

本文提出了一种实时的气温日值数据插补方法，方法利用一元线性回归方程将气温观测时间序列分解出拟合直线和残差曲线，并通过将二者再次重构实现气温序列的重组；方法给出了气温插补区的定义和插补区构成的充分条件，在满足条件的情况下，可以实现随机序列长度的插补需求；方法提出了采用动态时间规整衡量气温序列相似性的新模式，使得站点间的距离计算更加科学和精准，可以满足站点随机分布的插补需求。

本文利用山东省2021年的气温实况数据，对方法进行了双随机检验，检验结果表明：

(1)日平均气温、日最低气温和日最高气温三种气温日值要素均能够成功实现数据插补。

(2)在插补流程中采用DTW距离测度和二次插补的组合方法，插补效果优于目前常见的基于水平距离或海拔高度等地理临近关系的组合方法。

(3)方法对于地形存在一定的敏感性，在平原或丘陵地区的插补效果优于山地地区。

本文提出的基于时间序列的数据二次插补机制，对于解决日益增加的双随机特点的气象资料缺失问题，有着较为广阔的应用前景，也可为历史长序列气象数据的均一化订正提供一个良好的借鉴。

参考文献：

Della-Marta P M, Wanner H, 2006. A method of homogenizing the extremes and mean of daily temperature measurements[J]. J Climate, 19(17):4179-4197. doi:10.1175/JCLI3855.1.

Hansen J, Ruedy R, Sato M, et al., 2010. Global surface temperature change[J]. Rev Geophys, 48(4):RG4004. doi:10.1029/2010RG000345

余予,李俊,任芝花, 等,2012.标准序列法在日平均气温缺测数据插补中的应用[J].气象,38(9):1135-1139. Yu Y, Li J, Ren Z H, et al., 2012. Application of standardized method in estimating missing daily mean air temperature[J]. Meteor Mon, 38(9): 1135-1139. (in Chinese).

司鹏,郭军,赵煜飞, 等,2022.北京1841年以来均一化最高和最低气温日值序列的构建[J].气象学报, 80(1):136-152. Si P, Guo J, Zhao Y F, et al., 2022. New series of daily maximum and minimum temperature observations for Beijing, China since 1841[J]. Acta Meteorologica Sinica, 80(1):136-152.doi: 10.11676/qxxb2022.008. (in Chinese).

王海军,涂诗玉,陈正洪,2008.日气温数据缺测的插补方法试验与误差分析[J].气象,34(7):83-91. Wang H J, Tu S Y, Chen Z H, 2008. Interpolating method for missing data of daily air temperature and its error analysis[J]. Meteor Mon,34(7): 83-91. (in Chinese).

刘莹,师春香,王海军,等,2021.CLDAS气温数据在中国区域的适用性评估[J].大气科学学报,44(4):540-548. Liu Y, Shi C X, Wang H J, et al., 2021. Applicability assessment of CLDAS temperature data in China[J].Trans Atmos Sci,44(4):540-548.doi:10.13878/j.cnki.dqkxxb.20200819001. (in Chinese).

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

师春香,潘旸,谷军霞,等,2019.多源气象数据融合格点实况产品研制进展[J]. 气象学报,77(4):774-783. Shi C X, Pan Y, Gu J X, et al., 2019. A review of multi-source meteorological data fusion products[J]. Acta Meteorologica Sinica, 77(4):774-783.doi:10.11676/qxxb2019.043. (in Chinese)

龙柯吉,师春香,韩帅,等,2019.中国区域高分辨率温度实况融合格点分析产品质量评估[J]. 高原山地气象研究,39(3): 67-74. Long K J, Shi C X, Han S, et al., 2019. Quality assessment of high resolution temperature merged grid analysis product in China[J]. Plateau and Mountain Meteorology Research, 39(3): 67-74.doi:10.3969/j.issn. 1674-2184. 2019.03.011. (in Chinese).

中国气象局,2003.地面气象观测规范[M]. 北京:气象出版社,121. China Meteorological Administration, 2003. Specifications for surface meteorological observation [M].Beijing: China Meteorological Press, 121. (in Chinese).

俞剑蔚,李聪,蔡凝昊,等,2019.国家级格点实况分析产品在江苏地区的适用性评估分析[J].气象,45(9):1288-1298. Yu J W, Li C, Cai N H, et al., 2019. Applicability evaluation of the national gridded real-time observation datasets in Jiangsu province[J]. Meteor Mon, 45(9):1288-1298. doi:10.7519/j.issn.1000-0526.2019.09.009doi:10.7519/j.issn.1000-0526.2019.09.009. (in Chinese).

魏凤英,2007. 现代气候统计诊断与预测技术(第2版)[M]. 北京:气象出版社,37. Wei F Y, 2007. Modern climate statistical diagnosis and prediction technology(2nd edition) [M]. Beijing: China Meteorological Press, 37. (in Chinese).

Tormene T, Giorgino S, Quaglini M, 2008. Stefanelli matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation[J].Artificial Intelligence in Medicine,45(1),11-34.doi:10.1016/j.artmed.2008.11.007.

李正欣,张凤鸣,李克武,等.2014.一种支持DTW距离的多元时间序列索引结构[J].软件学报,25(3):560-575. Li Z X, Zhang F M, Li K W, et al., 2014.Index structure for multivariate time series under DTW distance metric[J].Journal of Software,25(3):560-575.doi:10.13328/j.cnki.jos.004410. (in Chinese).

闫宏宸,肖熙.2021.概率线性判别分析在语音命令词置信度判决中的应用[J].计算机系统应用,30(1):54-62. Yan H C, Xiao X, 2021. Application of probabilistic linear discriminant analysis in voice command confidence measures[J]. Computer Systems & Applications, 30(1):54-62. doi: 10.15888/j.cnki.csa.007732. (in Chinese).

周笑天,陈益玲,李芸,等.2022. 一种基于特征曲线的自动土壤水分观测数据异常值检测方法[J].中国农业气象,43(3):229-239. Zhou X T, Chen Y L, Li Y, et al., 2022. An outliers detection method for automatic soil moisture observation data based on characteristic curve[J]. Chinese Journal of Agrometeorology, 43(3):229-239. doi:10.3969/j.issn.1000-6362.2022.03.006. (in Chinese).

孙靖,程晓晓,黄小玉,2021.中国地面气象要素格点融合业务产品检验[J]. 高原气象,40(1):178-188. Sun J, Cheng G G, Huang X

Y, 2021. The verification of gridded surface meteorological elements merging product in China[J]. Plateau Meteorology, 40(1):178-188. doi:10.7522/j. issn.1000-0534.2019.00100. (in Chinese)

李超,唐千红,陈宇,等, 2017.多源数据融合系统LAPS的研究进展及其在实况数据服务中的应用[J]. 气象科技进展,7(2):32-38. Li C, Tang Q H, Chen Y, et al., 2017. An overview of progresses in LAPS and prospective applications in real time data service[J]. Advances in Meteorological Science and Technology, 7(2):32-38. doi:10.3969/j.issn.2095-1973.2017.02.005. (in Chinese).

潘旸,谷军霞,宇婧婧,等, 2018.中国区域高分辨率多源降水观测产品的融合方法试验[J]. 气象学报,76(5):755-766. Pan Y, Gu J X, Yu J J, et al., Test of merging methods for multi-source observed precipitation products at high resolution over China[J]. Acta Meteorologica Sinica, 76(5):755-766. doi:10.11676/qxxb2018.034. (in Chinese).

张璐,田向军,刘宣飞,等, 2017. 基于多重网格策略的NLS-3DVar资料融合方法及其在气温数据融合中的应用[J]. 气候与环境研究,22 (3): 271-288. Zhang L, Tian X J, Liu X F, et al., 2017. NLS-3DVar data fusion method based on multigrid implementation strategy and its application in temperature data fusion[J]. Climatic and Environmental Research, 22 (3): 271-288. doi:10.3878/j.issn.1006-9585.2016.16140. (in Chinese).

周笑天, 褚希, 姚志平, 2012.一种基于k-means聚类的实时气温动态质量控制方法[J]. 气象,38(10):1295-1300. Zhou X T,Chu X,Yao Z P, 2012. A dynamic method of quality control for real-time temperature measurements based on k-means cluster algorithm[J]. Meteor Mon, 38(10):1295-1300. doi:10.7519/j.issn.1000-0526.2012.10.016. (in Chinese)

王海军, 闫荞荞, 向芬, 等.2014.逐时气温质量控制中界限值检查算法的设计[J]. 高原气象, 33(6) : 1722-1729.Wang H J, Yan Q Q, Xiang F, et al., 2014. Algorithm design of quality control for hourly air temperature[J]. Plateau meteorology, 33(6) : 1722-1729.doi:10.7522 /j. issn.1000-0534.2014.00028. (in Chinese)

叶小岭,陈洋,杨帅,等,2019. 基于EEMD-CES的单站地面气温资料质量控制方法研究[J]. 大气科学学报,42(3):390-398. Ye X L, Chen Y, Yang S, et al., 2019. A quality control method of surface temperature observations based on the EEMD-CES algorithm for a single station[J]. Trans Atmos Sci, 42(3):390-398. doi: 10.13878/j.cnki.dqkxxb.20171205001. (in Chinese)

Twice Interpolation of Daily Temperature Based on DTW

ZHOU Xiaotian[1,2], ZHANG Qianru[1,2*], GUO Qingyan[3], CHEN Yiling[1,2], ZHOU Xuesong[1,2], LI Changjun[1,2], FENG Yong[1,2], ZHANG Ping[1,2]

1. *Key Laboratory for Meteorological Disaster Prevention and Mitigation of Shandong, Jinan 250031, China;*

2. *Shandong Meteorological Data Centre, Jinan 250031, China;*

3. *Fujian Meteorological Information Centre, Fuzhou 350001, China*

Abstract: As the most basic physical quantity for studying on climate evolution, the integrity and accuracy of daily temperature series are of great significance for climate analysis and assessment. In recent years, with the deployment of a large number of unmanned ground intensified automatic weather stations, missing data with double random characteristics such as random distribution of stations and random lengths of series, which poses significant obstacles

to climate analysis and operational applications. In view of the shortcomings of the existing methods for meteorological data interpolation, a new twice interpolation method of daily temperature data based on dynamic time warping (DTW) is proposed in this paper. The method adopts a real-time interpolation strategy, which mainly includes: (1) The method decomposes the temperature observation time series into a fitted straight line and a residual curved line by using univariate linear regression equation, and recomposes new temperature series by combining the two lines; (2) The method provides the definition and interpolation conditions of temperature interpolation areas; (3)The method proposes a new model for calculating the distance between stations by DTW . The collecting temperature data from Shandong Province in 2021 is used to test the method, and the test results show that the method can meet the interpolation needs of daily temperature data with double random characteristics, and the combination method of DTW distance and twice interpolation in the interpolation process can achieve a better effect than any of the other combination based on site geographical proximity relationships; the method is sensitive to terrain, and the interpolation effect in plain or hilly area is better than that in mountainous area.

Key words: Daily temperature; DTW; Recomposition; Twice interpolation