第 23 卷 第 1 期
2012 年 2 月

应 用 气 象 学 报
JOURNAL OF APPLIED METEOROLOGICAL SCIENCE

Vol. 23, No. 1
February 2012

王海军,刘莹. 综合一致性质量控制方法及其在气温中的应用. 应用气象学报,2012,23(1):69-76.

综合一致性质量控制方法及其在气温中的应用

王海军[1)2)*]　刘　莹[1)2)]

[1)](湖北省气象信息与技术保障中心，武汉 430074)

[2)](湖北省气象局气候变化中心，武汉 430074)

摘　要

由于历史逐日气温资料在气候分析、气候变化研究中的基础性作用,其数据质量状况日益受到关注。利用邻近参考站日平均气温、最高气温和最低气温资料及线性回归模型,设计了基于线性回归数据估计方法的质量检查算法,该算法同时包含了时间一致性和空间一致性两种检查方法。通过数据植入误差检测以及与单一空间回归检查方法的比较,该算法的错误数据检测性能较高,可检测出与正确日气温数据相差 3℃左右的可疑值。在该算法的基础上,研制了综合一致性数据质量控制方法,该方法具有以下特点:第一类错误发生率较低;保持了时间、内部和空间一致性的逻辑关系;参考了天气因素。因此,与一般的数据质量控制方法相比,综合一致性数据质量控制方法具有较高的错误数据检测性能。经过在华中区域湖北、湖南和河南三省 251 个站 1961—2009 年逐日气温资料的应用,取得了较好效果。各要素奇异值检出率平均气温为 0.001%,最高气温为 0.05%,最低气温为 0.04%。

关键词：逐日气温；线性回归；质量控制；综合一致性

引　言

近年来,我国学者非常关注气象观测数据质量,特别是地面气象数据,各种数据质量控制方法在气象业务、科研和服务中得到了一定应用。如任芝花等[1]、王海军等[2]、陶士伟等[3]、封秀燕等[4]在自动气象站的实时资料质量控制方面进行了研究,并设计了相应业务系统;廖捷等[5]对我国飞机观测气象资料进行了质量控制;任芝花等[6-8]对全球历史地面天气报资料数据集、1971—2000 年我国地面 700 多个基准基本站月气候资料进行了质量检查。国外在气象数据质量控制和评估方面比我国开展要早,近年来取得了不少成果[9-14]。

由于在气候诊断分析、气候变化研究中,日气温数据是非常重要的基本资料,如我国各区域气象中心正在实施的“区域气候变化评估报告编制”工作,就需要高质量的历史逐日气温数据。然而目前我国对于日气温数据的质量控制基本采用传统方法,使得隐含在历史气温数据中的错误没有得到系统和全面的检查,影响其应用,所以开展新型数据质量控制方法研究很有必要。为此本文探讨两方面内容:一是设计基于线性回归数据估计方法的质量检查算法(简称线性回归质量检查算法),二是在该检查算法的基础上,研制综合一致性数据质量控制方法,对华中区域三省的历史日气温数据进行质量控制,以提高数据质量。

1　数　据

本文研究所使用的数据为 1961—2009 年华中区域三省 251 个站日平均气温、最高气温、最低气温 3 个要素,并应用日降水量和日照时数作为参考要素,其中河南省有 105 个站,湖北省有 71 个站,湖南省有 75 个站。

上述资料来自各省气象档案馆的信息化资料。这些资料的数字化工作大体可分为 3 个阶段:①建站到20世纪90年代中期以前的资料,由人工首先

2011-04-06 收到，2011-11-22 收到再改稿。

资助项目:气候变化专项“华中区域气候变化评估报告编制”(CCFS-10-04),湖北省气象局基金课题(2009Y02)

* E-mail：whzxwhj@public.wh.hb.cn

从纸质报表输入到卡片或纸带，然后再转化成信息化资料；②20 世纪 90 年代中期到 2003 年前后，该阶段的资料由基层台站直接数字化后，按月上报到省级资料处理部门；③自动气象站阶段，从 21 世纪初开始我国陆续用自动站取代地面人工观测，其数据从采集开始就已数字化。所以本文所使用的资料绝大部分（自动站除外）都经过了人工观测、记录、抄录以及计算机录入、数据格式转换多个环节，所以不可避免地存在错误数据。这些数据虽已经过极值检查、内部一致性检查以及统计学检查等质量控制方法的检查，但对于与正确观测数据相差不大的可疑数据（如相差 3℃左右）上述方法检查效果大都不理想。

2 线性回归质量检查算法及其性能

气象资料质量控制方法中的空间一致性检查是较常用的方法。该方法的理论基础是气象要素在空间分布具有相关性，即空间距离较近的气象站点比距离较远的站点其特征值具有更大的相似性。空间一致性检查首先进行数据的空间插值，然后比较插值的估计值与观测值来实现数据的质量控制功能，其中使用较多的方法为空间回归检查方法[9-11,15-16]。

下面讨论的基于线性回归数据估计方法的质量检查算法（即线性回归质量检查算法），也包含空间一致性检查内容。该算法的思路是利用待检站与邻近站资料建立回归模型估计待检站数据，再通过观测值与估计值的差异，判断观测数据的质量。

2.1 线性回归数据估计方法

根据文献[17]，将被检站数据作为估计量，邻近站资料为因子，可建立多元线性回归模型：

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{e}, \tag{1}$$

式(1)中，$\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{e}$ 为向量，分别为被检站数据、待估计的参数和随机变量；$\boldsymbol{X}$ 为因子矩阵，表示邻近站资料。通过取样，对回归模型(1)进行估计，即可得到估计被检站气温的多元线性回归方程：

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p。 \tag{2}$$

式(2)中，$b_0,b_1,\cdots,b_p$ 为回归系数，是向量 $\boldsymbol{\beta}$ 的估计；$x_1,x_2,\cdots,x_p$ 为邻近站气温；$\hat{y}$ 为被检站估计值；p 为邻近站数。本文采用最小二乘法求解回归模型的系数。

求解回归模型的样本数据选取方法采用文献[18]的滑动优选法，即样本数据为被检日所在年前后若干年的同期资料。如估计 1981 年 5 月 1 日最高气温，则选择被检站以及邻近站 4 月 16 日—5 月 16 日（1979—1983 年）共 5 年每年 31 d 的最高气温数据作为样本数据（不含 1981 年 5 月 1 日数据），建立模型并求解回归系数，利用邻近站 1981 年 5 月 1 日的数据，通过式(2)，估计 1981 年 5 月 1 日的最高气温。为简化取样方法，当估计时间边界上的数据（如估计 1 月 1 日，则样本数据的日期要跨年）时，数据选样采用文献[18]的方法。其中邻近站是选取在距被检站半径为 150 km 范围内且拔海高度相差在 300 m 以内的 5 个气象站。

2.2 线性回归质量检查算法

根据文献[11]的原理，可通过计算日序为 i 的数据质控参数值来判断其质量状况，即

$$f_i=\frac{y_i-\hat{y}_i}{\sigma}。 \tag{3}$$

式(3)中，y_i 和 $\hat{y}_i$ 分别为日气温观测值和估计值；σ 为日气温估计误差的标准差，该标准差的统计时段为 1 个月（为被检日期所在月）；f_i 为质量控制参数（简称质控参数），其大小可表示观测值 y_i 的质量状况，显然 $|f_i|$ 越大，y_i 越可疑，甚至为错误数据。根据统计学原理，将质控参数值 $|f_i|\geqslant 3$ 的数据标注为奇异值。上述基于线性回归数据估计方法来计算数据的质控参数值，以判断数据质量的检查算法，称为线性回归质量检查算法。

2.3 线性回归质量检查算法性能评估

2.3.1 植入误差检测

本文采用文献[10-11]的方法，即通过植入误差，来评价线性回归质量检查算法的错误数据检测性能。该方法是将实际观测数据人为加上一定幅度的数值（即植入误差），然后采用质量检查算法对含有植入误差的数据进行质量检查，以检验不同幅度植入误差的检测能力。为此在华中三省选取 10 个站（其中湖南省、河南省各 3 个站，湖北省 4 个站）1961—2009 年的逐日气温资料，分别加上一定幅度的植入误差（1.2～3.0℃），然后利用式(3)计算 f_i，并统计 $|f_i|\geqslant 3$ 所占比例（表 1）。当植入误差取负值时，结果与表 1 数据基本相同，故对其不作分析。

由表 1 可知，当植入误差幅度为 3.0℃时，质控参数值为 3 的数据超过了 92%，其中平均气温超过了 99%。由此可见线性回归质量检查算法可检测日气温为 3℃以上的植入误差。

表 1　各种植入错误数据的质控参数值 $|f_i|\geqslant 3$ 所占比例(线性回归法)

Table 1　The percentage of seeded errors data for quality control parameter more than 3(using linear regression method)

植入误差/℃	平均气温/%	最高气温/%	最低气温/%
0.0	0.4	0.5	0.4
1.2	53.8	18.1	15.6
1.4	69.6	28.1	25.1
1.6	82.2	41.8	39.1
1.8	89.4	53.7	50.8
2.0	93.8	65.0	63.1
2.2	96.4	74.6	73.0
2.4	97.9	81.6	80.1
2.6	98.8	87.2	85.6
2.8	99.2	91.2	90.1
3.0	99.5	94.3	92.4

注:植入误差为 0.0℃时,表示实际观测数据 $|f_i|\geqslant 3$ 的比例。

2.3.2　与单一空间回归检验方法的比较

文献[9]比较了反距离加权插值法和空间回归检验方法,认为后者比前者性能更好。为此本文仅就线性回归质量检查算法与单一空间回归检验方法进行比较。

空间回归检验方法在利用邻近站进行插值时,并不是按照反距离加权插值法将最大权重赋给最近的台站,权重大小依据被插值站与邻近站的均方根误差大小来选取,具体算法见文献[11,16]。本文采用该方法进行数据估计时,从与被检站距离最近的 8 个站中挑选相关系数最大的 5 个站作为邻近参考站。

从统计结果来看(数据表略),虽然两种方法估计值的平均误差均很小,但单一空间回归检验方法的平均绝对误差和标准差远远大于线性回归检查算法,由此表明后者的数据估计精度明显高于前者。

由单一空间回归检验方法植入误差的检测情况(表 2)可看出,当植入误差的幅度为 3.0℃时,最高气温和最低气温只有 60%左右数据的质量控制参数值达到了 3,明显低于线性回归质量检查算法的比例(表 2 数据的统计台站和资料日期与表 1 相同)。

通过对两种方法的绝对平均误差、标准差和植入误差的检测情况比较,线性回归质量检查算法的错误数据检测性能明显优于单一空间回归检验方法,产生该现象的原因可能与它们的方法有关。虽然两者均使用了回归方法,但单一空间回归检验方法只是利用回归方法来确定与邻近站权重,然后根据该权重计算被检站数据大小[9-11];而线性回归质量检查算法是通过式(2)直接计算被检站的估计值。从这方面来说,线性回归质量检查算法是一种集时间一致性和空间一致性检查于一体的质量检查方法,而空间回归检验方法基本上还是较为单一的空间一致性检查方法。

表 2　各种植入错误数据的质量控制参数值 $|f_i|\geqslant 3$ 所占比例(空间回归法)

Table 2　The percentage of seeded errors data for quality control parameter more than 3(using spatial regression test)

植入误差/℃	平均气温/%	最高气温/%	最低气温/%
0.0	0.7	0.6	0.9
2.2	65.4	22.4	33.2
2.4	72.0	30.6	40.6
2.6	77.0	39.2	47.7
2.8	80.9	48.0	54.3
3.0	84.1	56.1	60.2
3.2	86.8	63.4	65.4
3.4	89.3	69.4	69.8
3.6	91.4	74.6	74.0
3.8	93.3	78.6	77.5
4.0	94.9	81.9	80.7

注:植入误差为 0.0℃时,表示实际观测数据 $|f_i|\geqslant 3$ 的比例。

3　综合一致性数据质量控制方法及检查步骤

线性回归质量检查法包含了时间和空间一致性检查两种方法,为提高质量控制效果,在此基础上增加内部一致性检查,并将该方法称为综合一致性数据质量控制方法。所增加的内部一致性检查是基于气温(日平均气温、最高气温、最低气温)、降水和日照时数等相关要素之间的变化规律来检查数据质量。因前文讨论的线性回归质量检查算法包含了时间和空间一致性方法,所以这里仅讨论新增的内部一致性检查方法。

3.1　气温要素之间的内部一致性检查

利用日平均气温、最高气温、最低气温 3 个要素相关性,即内部一致性来进一步检测数据质量。如某站某日的最高气温质控参数值 $|f_i|\geqslant 3$,表明其与周围站相比偏高。出现该现象有两种可能,一是该数据为奇异值,二是天气原因,即可能该站为晴天,而周围站为阴天或雨天。如为天气原因,一般来说日平均气温也偏高(和邻近站比),这样在检查最高气温时,同时参考日平均气温质控参数值,如同时偏高或偏低,一般认为是该站的天气和邻近站不一样造成,所检测的数据为有效值,即假设被检要素和参考要素同时为奇异值且同时偏大(或偏小)的可能性很小。该假设对于历史资料是合理可信的,因为历

史资料中逐日最高、最低气温和平均气温是由不同温度表测量得到，且观测时间也不一样，所以它们同时产生错误的概率非常小。如发生被检要素与参考要素质控参数值同时明显偏高(或偏低)现象，可认为是被检站天气与邻近站的差异造成，被检数据为有效数据，否则为奇异值。

下面以一个实例说明气温相关性的一致性检查情况(表 3)。湖北郧西站 1992 年 5 月 6 日最高气温质控参数值为 5.83，而同日的平均气温质控参数值为 4.51。如仅仅从最高气温质控参数值来判断，显然应被标注为奇异值，但如参考平均气温质控参数，则判定最高气温为有效数据。

表 3 1992 年 5 月 6 日郧西站及其邻近站日气象要素

Table 3 The meteorological elements of Yunxi Station and its neighboring stations on 6 May 1992

气象要素	郧西	邻近参考站				
		竹溪	郧县	竹山	房县	老河口
平均气温/℃	23.2	21.2	20.0	21.9	20.4	18.6
最高气温/℃	27.5	24.6	22.0	24.8	23.8	22.9
最低气温/℃	20.2	19.8	18.1	20.6	18.4	17.6
日照时数/h	2.3	0.0	0.0	0.0	0.0	0.0
降水量/mm	0.0	5.0	7.5	20.2	53.1	37.5

3.2 气温与降水量(日照时数)要素的内部一致性检查

如某站有降水，而邻近站均无降水，或某站无降水，而邻近站均有降水发生，这两种情况也会可能导致日气温被错误标注为奇异值。本文通过计算两个次序量 C_R，C_S 来配合气温的内部一致性检查。其中 C_R 表示待检站待检日的降水量大于邻近站同日降水量的站次数，C_S 表示待检站待检日的日照时数小于同日邻近站日照时数的站次数。这样可结合次序量 C_R，C_S 的大小实施降水量、日照时数等要素内部一致性检查。

3.3 综合一致性质量控制方法的步骤

3.3.1 逐日质控参数值和质量控制码计算

基于线性回归质量检查算法，利用式(3)计算气温的逐日质控参数值 f_i，依据表 4 规则标注各要素的初始质控码 F，质控码 F 的大小表示数据质量状况，其值越大，表明该数据为错误的可能性越大。

3.3.2 气温参考要素内部一致性检查

日最高气温、最低气温的参考要素为日平均气温，而日平均气温的参考要素为日最高气温、最低气温。对于日最高气温、最低气温，如同日的日平均气温质控参数值符号与其相同，则按照表 4 中参考要素的规则修正质控码。如符号不相同，则质控码不变。

对于日平均气温，同日的日最高气温或最低气温中如果有一个要素质控参数值符号与其相同，则按照表 4 中参考要素的规则执行(如果 2 个均同号，则取绝对值最大者)，如符号均相反，则平均气温质控码不变。

表 4 内部一致性质量控制方法标注规则

Table 4 The flagged rules of the internal consistency data check method

被检要素质控参数	参考要素质控参数	修正前质控码	修正后质控码
$\|f_i\|\geqslant 7$	$\|f_i\|\geqslant 3$	$F=4$	$F=1$
$5\leqslant\|f_i\|<7$	$2\leqslant\|f_i\|<3$	$F=3$	$F=1$
$3\leqslant\|f_i\|<5$	$1\leqslant\|f_i\|<2$	$F=2$	$F=1$
$\|f_i\|<3$	$\|f_i\|<1$	$F=0$	$F=0$

3.3.3 与降水量和日照时数要素内部一致性检查

对于最高气温、最低气温，当 $f_i<0$ 时，如 $C_R\geqslant 3$ 或 $C_S\geqslant 3$，则质控码 F 减 1。平均气温不参考降水量和日照时数要素，即对其不进行内部一致性检查。

3.3.4 质量状况的确定

经过 3.3.1～3.3.3 节所描述步骤，最后的质控码即为该数据的质控码。本文将数据质量状况分为 3 级，即错误、可疑和正确。其中 $F\geqslant 3$，表示该数据错误(奇异值)，$F=2$ 表示数据可疑，$F\leqslant 1$ 表示数据正确。

4 应用与讨论

4.1 一般质量控制方法的不足

在数理统计中定义了两种类型的错误，即如将正确数据标注成错误数据而拒绝，则称发生了第一类错误；如将错误数据标注成正确数据而接受，则称发生了第二类错误。在一般气象数据质量控制工作中，发生第一类错误的可能性较大。如仅采用线性回归质量检查算法，将最高气温 $|f_i|\geqslant 3$ 的数据标记为奇异值，就有 0.5% 左右的数据被标注(表 1)，这样在华中三省中就有两万多个日最高气温被标注，因该资料已经过基本质量检查，显然不可能包含如此多问题数据。当对部分标注的数据进行人工质量控制时，发现其中大部分数据并无质量问题，即第一类错误发生率较高。

在一般气象数据质量控制时,通常是按顺序提交各质量检查方法,最后综合各方法标注结果,决定数据的质控码。这样就失去了各方法的逻辑联系。如时间一致性检查某数据为奇异值(可疑,偏高),而空间一致性检查该数据也为奇异值(可疑,但偏低),则最后在综合确定其质量时,很有可能将其标注为错误数据(因该数据有两个主要的质量检查方法没有通过),但如将时间一致和空间一致性检查同时进行且参考奇异值偏离方向(偏高或偏低,即方法之间的逻辑联系),该数据就会被标注为有效数据。另外还存在一个问题,即目前许多数据质量控制方法很少考虑特殊天气和中小尺度天气对数据质量标注的影响。

综上所述,当前的一般数据质量控制方法存在 3 个方面的不足:第一类错误发生率较高;各种质量控制方法之间缺少逻辑性;大都基于气候原理和统计方法,很少考虑天气因素。

4.2　综合一致性质量控制方法的优点

4.2.1　第一类错误发生率较低

经统计,当最高气温质控参数值 $f_i \geqslant 3$(或 $f_i \leqslant -3$)时,该站同日参考要素平均气温的质控参数值 $f_i \geqslant 1$(或 $f_i \leqslant -1$)中的比例为 52%(即有 52%比例的日最高气温采用综合一致性质量控制方法中内部一致性检查后,可将标注为错误的数据降低为可疑数据乃至正确数据);而日最低气温参考要素达到相同质控参数值的比例为 68%,平均气温则为 85%。由此可见,仅气温相关性的内部一致性检查一项至少可减少第一类错误发生率一半以上,其中平均气温减少更多。如加上降水量、日照时数参考要素的内部一致性检查,综合一致性质量控制方法第一类错误发生率还会下降更多。

4.2.2　保持各种检查方法的逻辑关系

一般数据质量控制方法的质量控制码是一次标注,而综合一致性控制方法在质量控制过程中,依据数据发生异常的逻辑关系,对质量控制码的大小进行增加或减少;同时综合一致性质量控制方法是将时间一致性、空间一致性和内部一致性 3 种检查作为一个整体进行检查,在检查过程中,充分考虑了质量控制参数值大小与符号,这样就完整保留了各种方法的逻辑关系,不会出现一种方法检查某个数据偏大,另外一种方法检测出该数据偏小,而最后综合决策时将该数据标注为奇异值的情况发生。

4.2.3　参考天气因素

在气象数据质量控制工作中,常出现以下情况,即当发生一些中小尺度特殊天气现象时,导致很多无质量问题的数据被标注。Fiebrich 等[19]在对美国中尺度天气网的气象观测数据质量控制时,分析了特殊天气现象对数据质量标注的影响。本文在综合一致性质量控制方法中将降水量和日照时数作为日气温数据的参考要素进行内部一致性检查,某种程度上也是参考了部分特殊天气造成日气温空间分布不连续。

4.3　实例分析

经统计,湖北省云梦站 1985 年 12 月 22 日最高气温的质控参数值为12.78,平均气温的质控参数值为 0.42,可见云梦站该日最高气温与邻近参考站相比明显偏高,而日平均气温与邻近站相比差异不大(从表 5 中数据可知,云梦站该日最低气温与邻近站相比差异也不大)。所以从气象要素的内部一致性判断,该站该日最高气温可能为奇异值,且偏高。同时邻近站均无降水(表明邻近站气温不可能因雨降温而使该站的气温较邻近站偏高),由此表明,云梦站该日最高气温显著高于其他邻近台站的理由难以成立。

另外,从云梦站与邻近站该日前后的最高气温时间变化(图 1)来看,云梦站除该日明显偏高外,其他日期与其他站变化趋势相同。云梦站该日 4 个时次(02:00,08:00,14:00,20:00,均为北京时,下同)

表 5　1985 年 12 月 22 日云梦及邻近参考站气象要素

Table 5　The meteorological elements of Yunmeng Station and its neighboring stations on 22 Dec 1985

气象要素	云梦	邻近参考站				
		京山	安陆	应城	孝感	汉川
平均气温/℃	0.8	0.6	0.2	0.9	1.3	1.9
最高气温/℃	9.4	6.9	5.3	5.8	4.8	5.4
最低气温/℃	−1.8	−2.5	−1.4	−1.5	−1.4	−0.9
日照时数/h	0.0	1.6	0.0	1.2	0.0	0.0
降水量/mm	0.0	0.0	0.0	0.0	0.0	0.0

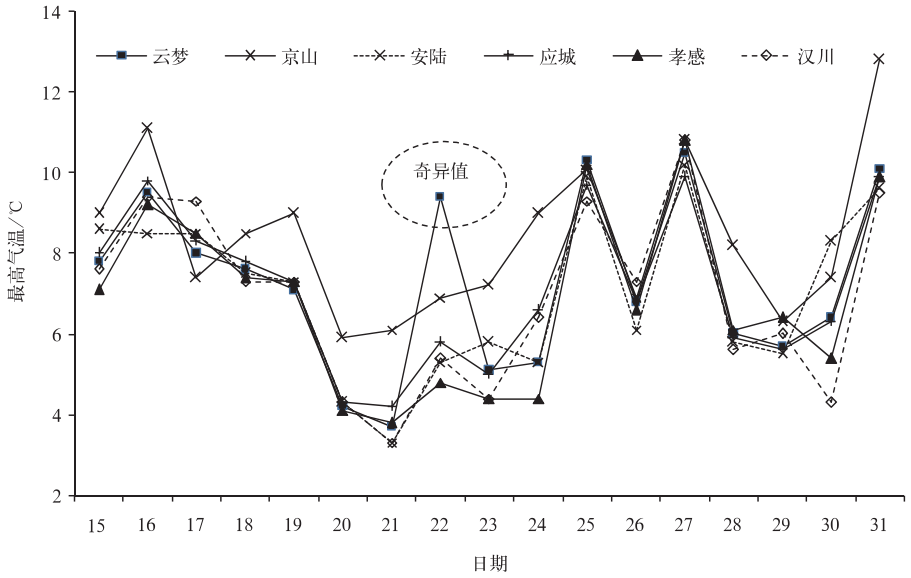

图 1 1985 年 12 月 15—31 日云梦及邻近站最高气温随时间变化

Fig. 1 The daily maximum temperature of Yunmeng Station and its neighboring stations during 15—31 Dec 1985

的正点气温分别为－0.2，－1.5，4.2，0.6℃，而该日日照时数为 0，所以日最高气温从 14:00 的 4.2℃上升到 9.4℃可能性较小。

综合上面的分析可知，云梦站该日最高气温 9.4℃为奇异值(其质控码 $F=4$)。该奇异值的产生很有可能是气温人工观测过程中较易出现的 5℃误读现象造成，即云梦站该日最高气温实际值可能为 4.4℃。

4.4 检查结果及分析

通过应用综合一致性质量控制方法的检查，华中三省日气温数据的奇异值检出率平均气温为 0.001%，最高气温为 0.05%，最低气温为 0.04%。经对奇异值分布分析发现，在标注为奇异值的数据中，最高气温的质控参数值为正的数据个数远远多于质控参数值为负的数量，而最低气温正好相反，即最高气温奇异值一般高于实际值，最低气温的奇异值一般低于实际值。分析表明，产生该现象的原因在于本文使用的数据前期经过基本质量控制，并修正了其中的错误数据。如最高气温发生与实际观测数据偏小的错误，将可能导致与逐小时的正点观测气温矛盾，从而前期在质量检查时，就会无法通过内部一致性检查，因此该错误数据得到了更正。但当最高气温发生偏高的错误(或最低气温发生偏低错误)时，内部一致性检查缺乏错误检测能力，故这些错误仍然被保留着。所以通过奇异值的分布特点可看出，综合一致性质量控制方法的确能检测出许多隐藏在数据集中的奇异值，从另一方面说明了该方法具有更好的错误数据检测性能。

5 小 结

本文主要任务是通过研制数据质量控制方法为华中区域气象中心研制一套高质量的日气温数据集，为区域气候变化评估报告提供数据支持。为此本文设计了基于线性回归数据估计方法的质量检查算法，并在该算法的基础上，研制了综合一致性数据质量控制方法，对 1961—2009 年华中区域三省 251 个站历史日气温资料进行了质量检查，取得了较好的应用效果：

1) 通过分析比较可知，基于线性回归数据估计方法的质量检查算法的错误数据检测性能较高，可检测出日气温与正确数据相差 3℃左右的可疑数据。

2) 综合一致性数据质量控制方法参考了天气因素，并在质量控制过程中保持了时间一致性、内部一致性和空间一致性的逻辑关系，与一般的数据质量控制方法相比，具有较高的错误数据检测性能。

本文在综合一致性数据质量控制方法中，尽管参考了中小尺度天气对数据奇异值标注影响，但天气因素对数据质量标注影响还需更深入研究，这也

是今后工作的方向。此外,由于在自动气象站中,日平均气温、最高气温、最低气温均为同一传感器的测量值,故综合一致性数据质量控制方法的气温相关性检查的检测效率会受到一定影响,可选择地温作为替代参考要素。

参 考 文 献

[1] 任芝花,赵平,张强,等.适用于全国自动站小时降水资料的质量控制方法.气象,2010,36(7):123-132.

[2] 王海军,杨志彪,杨代才,等.自动气象站实时资料自动质量控制方法及其应用.气象,2007,33(10):102-109.

[3] 陶士伟,仲跻芹,徐枝芳,等.地面自动站资料质量控制方案及应用.高原气象,2009,28(5):1202-1209.

[4] 封秀燕,何志军,王荷平,等.自动气象站实时资料质量控制开放式平台设计.应用气象学报,2010,21(4):506-512.

[5] 廖捷,熊安元.我国飞机观测气象资料概况及质量分析.应用气象学报,2010,21(2):206-213.

[6] 任芝花,刘小宁,杨文霞.极端异常气象资料的综合性质量控制与分析.气象学报,2005,63(4):526-533.

[7] 任芝花,许松,孙化南,等.全球地面天气报历史资料质量检查与分析.应用气象学报,2006,17(4):412-420.

[8] 任芝花,熊安元,邹凤玲.中国地面月气候资料质量控制方法的研究.应用气象学报,2007,18(4):516-523.

[9] Hubbard K G, You J S. Sensitivity analysis of quality assurance using the spatial regression approach—A case study of the maximum/minimum air temperature. *J Atmos Oceanic Technol*, 2005, 22: 1520-1530.

[10] Hubbard K G, Nathaniel B G, You J S, et al. An improved QC process for temperature in the daily cooperative weather observations. *J Atmos Oceanic Technol*, 2007, 24: 206-213.

[11] You J S, Kenneth G H. Quality control of weather data during extreme events. *J Atmos Oceanic Technol*, 2006, 23: 184-197.

[12] Durre I, Matthew J M, Byron E G, et al. Comprehensive automated quality assurance of daily surface observations. *J Appl Meteor Climatol*, 2010, 49: 1615-1633.

[13] Graybeal D Y, Arthur T D, Keith L E. Complex quality assurance of historical hourly surface airways meteorological data. *J Atmos Oceanic Technol*, 2004,21: 1156-1169.

[14] Graybeal D Y, Arthur T D, Keith L E. Improved quality assurance for historical hourly temperature and humidity: Development and application to environmental analysis. *J Appl Meteor*,2004,43:1722-1735.

[15] 任芝花,熊安元.地面自动站观测资料三级质量控制业务系统的研制.气象,2007,33(1):19-24.

[16] 刘小宁,鞠晓慧,范邵华.空间回归检验方法在气象资料质量检验中的应用.应用气象学报,2006,17(1):37-42.

[17] 黄嘉佑.气象统计分析与预报方法(第三版).北京:气象出版社,2004: 36-50.

[18] 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.气象,2008,34(7):83-91.

[19] Fiebrich C A, Kenneth C C. The impact of unique meteorological phenomena detected by the Oklahoma Mesonet and ARS Micronet on automated quality control. *Bull Amer Meteor Soc*, 2001, 82: 2173-2187.

Comprehensive Consistency Method of Data Quality Controlling with Its Application to Daily Temperature

Wang Haijun[1)2)] Liu Ying[1)2)]

[1)] (*Hubei Meteorological Information and Technological Support Center*, *Wuhan* 430074)

[2)] (*Climate Change Research Center*, *Hubei Provincial Meteorological Bureau*, *Wuhan* 430074)

Abstract

Due to the historical daily temperature data playing an important role in climate analysis and climate change research, the data quality is attached more importance. At present the daily temperature data are checked for quality control using the traditional methods in China, lacking a systematic and comprehensive method to pick up the outliers data hidden in the historical temperature data. These error data in the daily temperature affect data application, therefore, it's necessary to carry out the research of new quality control method.

Using linear regression model and historical daily temperature (average temperature, maximum temperature and minimum temperature) data of the neighbouring stations in the same period, a quality check algorithm based on linear regression estimation method is designed, which includes both time consistency check and spatial consistency check in quality control of meteorological observational data. To further enhance the detection performance of data quality check, a comprehensive consistency check method is developed based on this algorithm, which adds internal consistency check that refers the variation of related meteorological elements such as daily temperature (average temperature, maximum temperature and minimum temperature), precipitation and sunshine duration to check data quality.

Using the data seeded errors check test and compared with spatial regression test, the method of linear regression data quality control algorithm has higher error data check performance. The algorithm can detect suspicious data that is about 3℃ difference from the correct value on the temperature.

Through data quality control practices and analysis on historical data, the comprehensive consistency check method has the following advantages: The flagged rates of Type Ⅰ errors are lower, thus reducing false detection rate of that the correct data flagged as error data; the logical relationship are kept with time consistency, internal consistency, and spatial consistency in data quality control process, and these three methods of checking the consistency of data quality are as a whole at the same time; the weather factors are referred, thus reducing the impact on data quality of small-scale weather phenomena which can flag data incorrectly. Therefore, the method of comprehensive consistency data quality control, which compared to the traditional data quality control method, has higher error detection performance.

The algorithm achieves good progress on the applications of daily temperature data from 251 weather stations from 1961 to 2009 in Hubei, Hunan and Henan provinces. Detection of outliers in the average temperature is 0.001%, that in the maximum temperature is 0.05%, and that in the minimum temperature is 0.04%.

Key words: daily temperature; linear regression; data quality control; comprehensive consistency