空间回归检验方法在气象资料质量检验中的应用*

刘小宁　鞠晓慧　范邵华
(国家气象信息中心,北京 100081)

摘　要

该文详细介绍了空间回归检验方法,并使用 2003 年我国 671 站的逐日平均气温、最高气温、最低气温、平均水汽压、平均风速、平均 0 cm 地温、降水量资料,检验该方法在气象资料质量检验中的适用性。按区号将全国划分为 10 个区,利用该方法分别对各区 7 个要素进行了检验试验。结果表明:空间回归检验方法能够有效检验出可疑数据,适用于对单一要素的检验;对降水、风速等空间变化比较大的要素,该方法有比较好的检验效果;应用该方法计算时,在不同地区、不同要素之间存在差异;当固定出错比率时,各区应该选择不同的 f 值。与一般空间检验方法相同,该方法也与地理环境、周边台站分布有关,并受台站密度的影响。

关键词:空间回归检验方法;气象资料;适用性

引　言

气象观测资料质量控制的重要性已经为所有使用气象资料的科学家所公认。地面气象观测记录必须具有代表性、准确性和比较性[1]。在"全球气候观测系统资料与信息管理计划"中特别提出,为 GCOS 所建立的数据库系统中的数据必须经过严格的质量控制,使数据质量符合标准,以保证使用此类数据的研究人员能够接受[2]。

随着气候研究的深入,各国都进一步加强了对气象资料的质量控制(QC)研究。近年来资料质量控制技术有了新的进展,同时,质量评估(QA)也已经发展和应用到各国资料部门中。在质量控制的同时,对资料进行评估是必要的。尤其对进行质量控制的数据管理工作人员,因为评估可以使其更好地决定是否使用某方法,并且对检验出的结果有一个科学的认识。当然,如果将评估结果提供给资料的使用者,则可以使资料使用人员更好的利用资料。因此,在国外,近年来除了质量控制(QC)技术发展很快外,质量评估(QA)也更加引起重视[3-7]。

传统的质量控制(QC)主要根据气象学、天气学、气候学原理,以气象要素的时间、空间变化规律和各要素间相互联系的规律为线索,分析气象资料是否合理。其方法包括:范围检查、极值检查、内部一致性检查、空间一致性检查、气象学公式检查、统计学检查、均一性检查[8]。这些方法被普遍应用到地面气象资料的质量控制中,而且方法的综合应用有着很好的效果。随着自动站资料空间密度的增加,尤其是高密度降水、气温、风资料的使用,更需要对单要素的资料质量进行有效的控制。但是如果是单要素,或者是新建的测站或测点(如在我国大城市或乡镇密布的自动雨量站),由于缺乏历史资料或要素之间的相互比较,对其观测数据的质量控制就只有使用空间一致性检查比较合适。有许多进行空间插值估计的方法[9],其取决于气象变量、地理环境、台站的空间分布等,如:正态比例法、反距离权重法、内插法、单一最优估计、回归法等。对于空间检验,我国资料检验部门还很少使用。因此,本文介绍一种国外使用的"空间回归检验方法"[10],并应用该方法对 2003 年地面观测的主要气象要素进行检验,评估该方法在我国各地的适用性。期望对该方法的使用提供试验依据,促进对单要素高密度资料检验技术的提高。

1　资料和方法

1.1　资　料

为了检验方法的适用性,使用 2003 年我国 671 站的逐日平均气温、最高气温、最低气温、平均水汽

* 2005-01-14 收到,2005-04-07 收到再改稿。

压、平均风速、平均 0 cm 地温(以下简称 0 cm 地温)、降水量。同时为了便于分析该方法在各地区的适用性,按区号将全国台站划分为 10 个区。

50 区包括 45° N 以北的黑龙江及内蒙古北部;51 区为 35° N 以北、92° E 以西的新疆大部地区;52 区位于 35° N 以北、92° ~ 105° E 的地区,主要包括内蒙古西部、甘肃、青海;53 区位于 35° N 以北、105° ~ 115° E 的地区,主要包括陕西、山西、河北西部;54 区位于我国 35° ~ 45° N,115° E 以东地区,主要包括辽宁、吉林、内蒙古东部、河北东部、京津、山东地区;55 区位于我国 35° N 以南、92° E 以西的青藏高原地区;56 区位于 35° N 以南、92° ~ 105° E 地区,包括青藏高原东部、四川西部、云南大部;57 区位于 25° ~ 35° N,105° ~ 115° E 地区,主要包括四川东部、河南南部、湖北、湖南、广西等地;58 区位于 25° ~ 35° N,115° E 以东地区,主要包括安徽、江苏、江西、福建;59 区位于 25° N 以南、105° E 以东地区,主要包括广东、广西。

1.2 方 法

空间回归检验方法的主要思路是以参考站的观测值来判断被检验站的观测值是否在一定可接受的范围之内。首先选择距被检验站最近的一批站(本次选择距被检验站正负两个经纬度范围)。在这一批站中,以参考站和被检验站的一元回归方程的均方根误差为依据,具有最小均方根误差的 5 个站为确定的参考站。给均方根误差小的参考站以大的权重来计算被检验站观测值的加权估计值和估计值的加权标准差。然后用加权估计值和估计值的加权标准差确定被检验站实测值的范围,如果超出范围,则认为数据可疑。

具体计算步骤如下:

① 相对于被检验站,选择距被检验站最近的一批站。

② 每个站全月逐日数据作为一个序列,逐对计算参考站与被检验站之间的相关系数,并进行相关系数的显著性检验,对于没有通过显著性检验的台站,从参考站中剔除。

如果参考站少于 5 个,则该被检验站不再做进一步的空间检查。

③ 逐对建立被检验站与各参考站的回归方程:

$$\hat{x_i} = a + by_i \tag{1}$$

式(1)中,$\hat{x_i}$ 为被检验站逐日要素估计值,y_i 为参考站逐日要素值。

④ 确定 5 个参考站。求每对被检验站与参考站建立的回归方程的均方根误差:

$$s^2 = \frac{1}{n - m - 1}\sum_{i=1}^{n}(x_i - \hat{x_i})^2 \tag{2}$$

式(2)中,s 为均方根误差,x_i 为被检验站某日的实测值,$\hat{x_i}$ 为被检验站该日的估计值,因为是一元回归方程,$m = 1$,n 为日数。

对每一对回归方程的均方根误差由小到大排序后,找到前 5 个最小值。即确定有最小均方根误差的 5 个参考站。

⑤求加权估计值。计算被检验站的加权估计值 x':

$$x' = \sqrt{\sum_{i=1}^{n}\hat{x_i}^2\bar{s_i}^2}\Big/\sqrt{\sum_{i=1}^{n}\bar{s_i}^2} \tag{3}$$

式(3)中,$\hat{x_i}$ 为相应 s_i 的该日估计值,n 为站数(由于挑选的是 5 个站,所以 $n = 5$),s_i 为所选 5 站的均方根误差。

⑥ 求估计值的加权标准差 s':

$$\frac{1}{s'^2} = \frac{1}{n}\sum_{i=1}^{n}\bar{s_i}^2$$

$$s' = \sqrt{\frac{n}{\sum_{i=1}^{n}\bar{s_i}^2}} \tag{4}$$

⑦ 检验被检验站实测值,如果超出范围,认为数据可疑。

$$x' - fs' \leqslant x \leqslant x' + fs' \tag{5}$$

式(5)中,f 为质量控制参数,简称质控参数。

对每个被检验值 x 重复 ⑤~ ⑦步,逐值检查。

2 对空间回归检验方法的评估

根据概率统计学[11],我们知道,当检验一个假设时,总希望得到一个正确的判断,但是实际上完全做到是不可能的。无论是接受还是拒绝一个假设,所做的判断都有可能是错误的。其中有两种类型的错误,其一类错误在数理统计中称为“第一类错误”即“把真当假”。在做质量控制的统计检验时,如果数据正确而被拒绝,则产生“第一类错误”。其二是“以假充真”,即数据是错误的而被接受,产生“第二类错误”。

在空间回归检验方法的使用中,为了减少“第一类错误”,可以采取扩大判别范围的方法,在式(5)中扩大 f 值的数值,但同时将增多“第二类错误”,即扩大判别范围后,增多了可能是错误数据被认为正确

而接受。QA 研究的必要性就是在减少“第二类错误”而不增加“第一类错误”之间找到平衡。一般情况是,控制“第一类错误”的概率,在此基础上选择使“第二类错误”达到最小的检验方法。在质量控制中,如何选择判断的参数,使数据经过检验后,数据的质量达到一个适当的标准是重要的。

根据空间回归检验方法,我们可以看到,式(5)中,如果被检验数据 x 在给定范围之内,则数据通过检验;从式(5)可以看出,使用连续大的 f 值,可以减少“第一类错误”。因此,空间回归检验方法可以事先选择一个检测出的错误数据的比例,其中的 f 值是一个变化的范围,而不仅仅是单一的 f 值。应用该方法,在选择统一出错率的基础上,通过 f 值的变化,使质量控制后的数据评估有一个可以接受的统一标准。因此,可以在不同的气象站选择不同的 f 值,这是该方法的第一个特点。

该方法的第二个特点是,参考站不一定是与被检验站距离最近的站。参考台站的选取首先选距离近的站,在此基础上再根据被检验站与参考站建立的回归方程的均方根误差为权重,也就是使用了拟合最好的站来做为参考站,用拟合最好的几个站来估计被检验站的数据。所以,该方法的实质是:在被检验站邻近的一定空间范围内,寻找与它相关性较好的站做为参考站。这是因为一元回归系数也就是相关系数,相关系数高,相关较好,则方差也较小。虽然“空间回归检验方法”是一个考虑“空间”的检验方法,但不象空间检验中的“反距离权重法”仅以距离做为权重估计被检验站的数据。因此,空间回归检验方法比只考虑“距离”的“反距离权重法”有着更好的检验效果[12]。

3　检验结果及分析

表 1 列出了各区在不同 f 值下检验出的日平均气温平均出错率。应该说明的是:①表 1 中的平均值是由所有站检验出的认为是出错(可疑)的样本数除以全部被检验的样本数,不是各区出错率的平均值;②各区检验的样本数是有差别的,由于计算时必须满足参考站数不低于 5 个,在台站稀疏的区,检验样本就少,如 55 区的检验样本少。但是每个区出错率计算方法均为检验出的出错样本数除以检验样本数。

表 1　不同 f 值情况下,各区日平均气温平均出错率　　%

f	50 区	51 区	52 区	53 区	54 区	55 区	56 区	57 区	58 区	59 区	平均值
1	17.5647	18.6285	20.0925	18.0292	17.1273	21.1666	19.9506	15.9782	16.1059	16.8111	17.6672
2	0.8524	0.9968	1.9293	0.7203	0.6978	1.3430	1.2827	0.8113	0.7501	1.0043	0.9655
3	0.0122	0.0283	0.6564	0.0000	0.0122	0.0200	0.0466	0.0080	0.0328	0.0275	0.0703
4	0.0000	0.0000	0.3996	0.0000	0.0000	0.0000	0.0107	0.0000	0.0000	0.0000	0.0326
5	0.0000	0.0000	0.2797	0.0000	0.0000	0.0000	0.0072	0.0000	0.0000	0.0000	0.0228
6	0.0000	0.0000	0.2055	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0161
7	0.0000	0.0000	0.1884	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0147
8	0.0000	0.0000	0.1712	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0134
9	0.0000	0.0000	0.1712	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0134

从表 1 可以看出,各区气温要素均是随着 f 值的扩大,判为可疑数据的比率逐步减少,同时,各区减少的速度相差比较小。当 $f=3$ 时,平均气温全国平均出错率为 0.07 %;除 52 区外,各区均已经达到万分之几或更低的量级。当 $f=4$ 时,除 56 区和 52 区外,出错率均为 0.0。如果出错率选择在 0.01 %,可选择 f 值在 3 ~ 4 之间。当 $f=9$ 时,52 区出错率在 0.1712 %,其余区没有检查出错,经查,52 区的 52645 站(青海省野牛沟)6 月全月数据判错。$f=9$ 是一个很大的容错范围,数据仍然判为出错,因此,可肯定判定该站月数据有错。如去掉 52645 站,经检验,当 $f=7$ 时,52 区出错率在为 0.01 %。因此,一般地,52 区在出错率为 0.01 %时,可选择 $f=7$。

从以上计算结果可以看出,当 f 值固定时,各区的出错率相差比较大。当 $f=3$ 时,57 区出错率达到十万分之几的量级、52 区为千分之几的量级、其余区在万分之几的量级。因此,检验气温时,应该根据一般出错率经验,在我国不同地区选择不同的 f 值,使检验出的出错率控制在一定的比例内。也就是说,如果固定出错率,则不同地区应该选择不同的 f 值。

从表 2 可以看出,各区降水量也随着 f 值的扩大,判为错误数据的比率逐渐减少。但是,与气温要

素不同,各区减少的速度相差比较大。在检验降水中,当 $f=3$ 时,平均出错率为 5.42%,这一比例高于一般的经验出错率,因此,不能取 $f=3$。当 $f=5$ 时,平均出错率为 0.04%,除 52 区和 54 区外,均在万分之一以下。这一比例符合一般的经验出错率。因此,当使用该方法检验降水量时,取 $f=5$ 是合适的。当 $f=9$ 时,52 区和 54 区仍错,经查,52576 站 6 月 25 日和 54471 站 2 月 25 日数据判为有错,因此判断这些数据有错。如果不检验 54471 站,则当 $f=5$ 时,54 区出错率为 0。说明一般情况下,54 区选取 $f=5$ 比较合适。

表 2 不同 f 值情况下,各区降水量的平均出错率 %

f	50 区	51 区	52 区	53 区	54 区	55 区	56 区	57 区	58 区	59 区	平均值
1	47.6802	54.8193	49.5974	51.6478	47.3938	44.8718	47.4317	42.7840	41.4355	41.9052	44.7702
2	19.3485	34.9398	23.3494	22.5283	19.9807	6.4103	19.6721	16.6932	14.6526	16.9499	17.9576
3	7.6999	15.6627	8.8567	6.7388	7.3359	2.5641	6.4481	4.3809	3.3249	5.2773	5.4278
4	2.2705	3.6145	1.7713	1.4265	1.5444	0.0000	1.5301	0.6183	0.6879	0.7156	1.0499
5	0.0000	0.0000	0.6441	0.0000	0.0322	0.0000	0.0000	0.0177	0.0000	0.0000	0.0445
6	0.0000	0.0000	0.4831	0.0000	0.0322	0.0000	0.0000	0.0000	0.0000	0.0000	0.0198
7	0.0000	0.0000	0.3221	0.0000	0.0322	0.0000	0.0000	0.0000	0.0000	0.0000	0.0148
8	0.0000	0.0000	0.1610	0.0000	0.0322	0.0000	0.0000	0.0000	0.0000	0.0000	0.0099
9	0.0000	0.0000	0.1610	0.0000	0.0322	0.0000	0.0000	0.0000	0.0000	0.0000	0.0099

降水量是空间变率比较大的要素,空间检验的许多方法对降水要素的检验均不理想。而从以上试验及分析可以看出,该方法对降水的检验有比较好的效果,这也是该方法的一个优势。

从表 3 可以看到,当 $f=3$,平均出错率为 0.06%;当 $f=4$,平均出错率为 0.0008%。因此,在检验平均风速时,可以取 $f=4$。由于 55 区各站风速的相关差,造成参加检验的站少,表 3 中 55 区参加检验的样本少,出错比例下降比较快。由此也可以看出,对于风速这一空间变率大的要素,使用该方法检验,也有比较好的效果。

表 3 不同 f 值情况下,各区平均风速的平均出错率 %

f	50 区	51 区	52 区	53 区	54 区	55 区	56 区	57 区	58 区	59 区	平均值
1	23.3027	21.7479	23.0084	23.2465	22.5791	24.2366	24.4042	22.9665	22.6688	22.8883	22.9062
2	1.8003	2.0844	2.3323	1.8504	1.7234	1.8130	2.1116	2.1029	1.7651	1.7956	1.8676
3	0.0506	0.1507	0.1048	0.0468	0.0597	0.0000	0.0603	0.0876	0.0706	0.0489	0.0660
4	0.0000	0.0000	0.0000	0.0067	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0008
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

同样,应用空间回归检验方法检验了最高气温、最低气温、0 cm 地温、水汽压(表略)。结果表明:检验最高气温,当 $f=3$ 时,平均出错率为0.0655%,而当 $f=4$ 时,除 52 区有错外,均未再检验出错。因此,f 取 3~4 之间比较合适。经查,52 区是 52645 站 6 月资料有错。检验最低气温,当 $f=3$ 时,平均出错率为 0.05291%,而当 $f=5$ 时,除 52 区有错外,均未再检验出错。因此,f 取 3~4 之间比较合适。经查,52 区也是 52645 站 6 月数据出错。检验 0 cm 地温,当 $f=4$ 时,平均出错率为 0.05980%,而当 $f=5$ 时,平均出错率为 0.01483%。因此,f 取 4~5 之间比较合适。检验水汽压,当 $f=3$ 时,平均出错率为0.05045%,而当 $f=4$ 时,平均出错率为 0.01045%。因此,f 取 3~4 之间比较合适。

根据以上各要素的计算,综合出各要素不同 f 值的全国平均出错率,见表 4。

从表 4 可以看出,当 $f=3$ 时,除降水量、0 cm 地温外,各要素的出错率在 0.07%以下,即万分之几的量级。当 $f=4$ 时,0 cm 地温出错率为 0.05%;当 $f=5$ 时,降水量出错率达到 0.04%。这一结果可以作为各要素全国范围内选择 f 值的参考。即当选择出错率为万分之几的量级时,对降水量取 $f=5$,

表 4　不同 f 值情况下,各要素的平均出错率　　%

f	平均气温	最高气温	最低气温	降水量	风速	0 cm 地温	水汽压
1	17.6672	17.0123	19.1767	44.7702	22.9062	20.7300	17.8138
2	0.9655	1.1724	1.2587	17.9576	1.8676	1.7949	1.1431
3	0.0703	0.0655	0.0529	5.4278	0.0660	0.2220	0.0504
4	0.0326	0.0215	0.0077	1.0499	0.0008	0.0598	0.0104
5	0.0228	0.0170	0.0009	0.0445	0.0000	0.0148	0.0022
6	0.0161	0.0157	0.0000	0.0198	0.0000	0.0027	0.0004
7	0.0147	0.0134	0.0000	0.0148	0.0000	0.0004	0.0000
8	0.0134	0.0121	0.0000	0.0099	0.0000	0.0000	0.0000
9	0.0134	0.0080	0.0000	0.0099	0.0000	0.0000	0.0000

0 cm地温取 $f=4$;其余要素取 $f=3$。对于某一区,f 取值可与全国平均取值不同。同样,一个区的不同要素之间也有差别(表 5)。从表 5 可看出,具体到 50 区,选择出错率为 0.01 %时,对降水量要素取 $f=5$,对平均风速 $f=4$,其余要素 $f=3$。因此,在检验各区的单要素资料时,可以以表 5 为例,制作各区在固定出错率值后 f 的判断值。

表 5　不同 f 值情况下,50 区各要素的平均出错率　　%

f	平均气温	最高气温	最低气温	降水量	风速	0 cm 地温	水汽压
1	17.5647	16.9330	19.7017	47.6802	23.3027	20.4527	16.8402
2	0.8524	1.0553	0.9498	19.3485	1.8003	1.1410	1.0472
3	0.0122	0.0183	0.0183	7.6999	0.0506	0.0061	0.0183
4	0.0000	0.0000	0.0000	2.2705	0.0000	0.0000	0.0061
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

从以上分析可以看出,当使用该方法进行质量检验时,不同要素在不同地区检验的出错率不同,为了在实际检验中应用该方法对不同要素的检验,根据试验结果总结出表 6 供参考。在实际检验中,通常要求统一出错率标准,表中选择了出错率为 0.01 %。

表 6　当出错率为 0.01 %时,各区各要素的 f 值

	平均气温	最高气温	最低气温	水汽压	平均风速	0 cm 地温	降水量
50 区	3	3	3	3	4	3	5
51 区	4	4	4	4	4	6	5
52 区	7	7	5	6	4	5	6
53 区	3	4	4	4	5	4	5
54 区	3	3	4	4	4	4	5
55 区	4	4	4	4	3	4	4
56 区	4	4	4	4	4	7	5
57 区	3	4	4	4	4	5	6
58 区	4	4	4	3	4	4	5
59 区	4	4	4	4	4	4	5

从表 6 中可以看出,各地区各要素的 f 值有比较大的差别。当固定出错率时,各区选择的 f 值不同。如气温选择 0.01 %的出错率,50,53,54,57 区选择 $f=3$;52 区选择 $f=7$;而其余区选择 $f=4$。在理论上,出错率是在一定可信度的标准下被判断为错误的数据。在实际应用中,是一种可疑数据。为了避免“以真当假”的情况,还必须针对具体检验出的数据进行人工判断,以确定是“可疑”还是“错误”。如果经过人工判断,“可疑”的数据是“正确”的结果比较多,则应该调整 f 值,以降低出错率。

从总体来看,51,52,56 区的各要素 f 值合计大于其他区,这几个区的站点位于我国西部或西北部,地处我国边境或青藏高原,台站分布稀疏、地形复杂。55 区比较特殊,虽然也位于青藏高原,但由于

台站稀疏,参加检验的样本数少,通过检验的站结果比较好,所以各要素 f 值合计值也不大。这说明,与一般空间检验方法相同,该方法也与地理环境、周边台站分布有关,受到周边台站密度的影响。

4 小 结

1) 空间回归检验方法能够有效检验出可疑的数据,适用于对单一要素的检验。尤其对降水、风速等空间变化比较大的要素,该方法有比较好的检验效果。

2) 应用该方法计算时,在不同的地区、不同要素之间存在差异。随着 f 值的扩大,各区各要素判为错误数据的比率均逐渐减少。虽然如此,各要素减少的幅度有比较大的差别。降水量要素各区减少的速度差别比较大;其余要素减少的速度各区相差比较小。

3) 当固定出错率时,各区应该选择不同的 f 值。与一般空间检验方法相同,该方法也与地理环境、周边台站分布有关,受到台站分布密度的影响。

参 考 文 献

[1] 中国气象局. 地面气象观测规范. 北京:气象出版社,2003.

[2] 全球气候观测系统中国委员会办公室. 全球气候观测系统中国委员会成立大会暨委员会专家组第一次全体会议文集. 北京:气象出版社,1997:67.

[3] WMO. Guidelines on Quality Control Procedures for Data from Automatic Weather Stations. CBS/OPAG-IOS/ET AWS-3/DOC.4(1), 2004.

[4] Rudel E. Report and Review about Data Processing and Quality Control Procedures Involved in the Conversion of Manually Operated Station to Automatically Operated Station. World Climate Programme: zData and Monitoring No.31, WMO-TD No. 833, 1997.

[5] WMO. Automated Weather Stations for Applications in Agriculture and Water Resources Management: Current Use and Future Perspectives. Proceedings of an International Workshop. 6-10 March 2000, Lincoln, Nebraska, USA, WMO/TD No. 1074, 2001.

[6] Mark A Shafer, Christopher A Fiebrich, Derek S Arndt. Quality assurance procedures in the oklahoma mesonetwork. *J Atmos Ocean Technol*, 2000, 17: 474-494.

[7] Hubbard K G, Goddard S, Sorensen W D, et al. Performance of quality assurance procedures for an applied climate information system. *J Atmos Ocean Technol*, 2005, 22(1): 105-112.

[8] 吴忠义,马尚风译. 世界气候资料计划文件汇集. 北京:气象出版社,1990:106-109.

[9] Jon K Eischeid, C Bruce Baker, Thomas R Karl, et al. The quality of long-term climatological data using objective data analysis. *J Appl Meteorol*, 1995, 34: 2787-2795.

[10] You Jinsheng, Kenneth G Hubbard, Steve Goddard. Comparison of air temperature estimates from spatial regression and inverse distance method. http://www.hprcc.unl.edu/manuscripts/.

[11] 屠其璞,王俊德,丁裕国,等. 气象应用概率统计学. 北京:气象出版社,1984:183-186.

[12] Song Feng, Hu Qi, Qian Weihong. Quality control of daily meteorological data in China, 1951—2000: a new dataset. *Int J Climatol*, 2004, 24: 853-870.

A Research on the Applicability of Spatial Regression Test in Meteorological Datasets

Liu Xiaoning　Ju Xiaohui　Fan Shaohua
(*National Meteorological Information Center, Beijing* 100081)

Abstract

With the rapid growth of the AWS spatial distribution density, it is more rational to use spatial consistency checks in quality control of meteorological observations of some newly built stations and single element observing stations(i.e. the automatic rainfall stations densely covered all over our country). For that under these situations traditional historical comparative study and different elements comparing are difficult to do for lacking of data.

A new spatial regression checking method used abroad is introduced in detail and applied to spatial checking of some basic surface observing meteorological elements of China for the year of 2003 in order to evaluate the applicability of this approach in China. The method is designed for identification of suspected observing values among neighboring observations. First, some neighboring stations are selected by distance. Second, the root mean square (RMS) errors of the univariate regression equations which are established basing on the examined station observation and neighboring observations are calculated and five reference stations are determined by minimizing root mean square errors. The five reference stations are weighted differentially. Stations with smaller RMS errors get more weighting points. Then, the weighting estimate values and their weighting standard errors of the examined station are computed and used to determine the data range. Data not in this range would be flagged suspected.

The spatial checking tests are conducted on 7 basic meteorological elements including daily mean temperature, maximum temperature, minimum temperature, mean vapor pressure, mean wind speed, mean surface temperature and precipitation. The data are obtained from 671 weather stations all over China, and to get more reasonable results the data are divided into 10 districts according to their station designator.

Results show that this method works well in identifying errors of single meteorological element especially to the elements with larger spatial variation such as precipitation and wind speed. It should be noticed that there are some differences when applied this method to different areas and different elements. To get a fixed error rate, the values of f should be selected according to the corresponding districts. Finally, same as other spatial checking methods, geographical environment and distribution of neighboring weather stations should be concerned necessarily as influence factors. The approach performs poorer under the condition of sparse station density.

Key words: spatial regression test; meteorological datasets; applicability