山区农业气候区划中年平均气温空间序列的正态性研究

康锡言　高建华

(河北省气象局气象科学研究所,石家庄 050021)

摘要　选取河北省西部地区 30 个国家基本站 30 年年平均气温作为样本,通过对总体样本、分区样本的偏度系数、峰度系数的分析,研究了各站点组成的空间序列,年平均气温以及地理参数(海拔高度、经度、纬度)的正态性。研究发现:总体样本的年平均气温、海拔高度不遵从正态分布,而经度、纬度能较好的遵从正态分布;按照经纬度跨度不大于 1.5°,尽量多选取山区站;或既要多选山区站,又要使同一分区的站点具有相同的气候特点两种原则,对站点进行分区,能使分区样本的年平均气温和地理参数均遵从正态分布。

关键词　正态　偏度系数　峰度系数　气候区划

引言

农业气候资源由于纬度、海陆分布以及地势地貌与下垫面的特征不同,造成大范围的光、热、水资源在空间上有明显的区域差异。气象站点的数据虽然能部分地反映区域内农业气候资源的分布情况,但由于这些资料来源于气象台站的观测网,而观测点大部分在海拔较低的平坦地区,不能反映区域立体的农业气候资源分布,为了客观地反映区域内农业气候资源的立体分布特征,必须建立空间分析模型[1]。由于某点气温的变化,首先决定于该点所接收的太阳辐射,而接收辐射的多少直接受该点的纬度、经度、拔海高度等地理参数影响,这样就需要用地理参数与气温建立多元回归模型,分析计算区域内的气温分布状况。

在使用最小二乘估计建立多元回归模型时,大多是基于观测数据是多元正态分布的一个样本的假设[2],因此在回归分析中,对研究的变量都应进行正态分布检验。只有变量遵从正态分布,其数学期望才和众数一致。例如回归方程预报中,预报值是因子出现条件下预报量的期望值,预报量遵从正态分布,期望值自然是出现概率较大的值,即预报期望值恰好是最可能的出现值。如果预报量遵从的是偏态分布,预报得到的期望值不能对应现象出现的是较大概率值。因此研究变量的正态分布,是建立较准确的多元回归模型的基础,是十分必要的。

年平均气温为全年热量状况的总标志,而热量是作物生育不可缺少的环境条件之一,较准确地模拟山区年平均气温的分布状况,对调整作物布局,改革农业生产结构,具有重要的指导意义。为此,本文利用河北省西部地区 30 年的年平均气温资料,研究了年平均气温及地理参数(海拔高度、经度、纬度)的正态性;同时,对应用遵从和不遵从正态分布变量建立的回归模型的残差进行了分析。

1　资料与方法

选取河北省西部山区和山麓平原地区 30 个国家基本站(见表 1),1971 ~ 2000 年 30 年年平均气温,以及各站的经度、纬度、海拔高度组成的对象场作为研究对象,对 30 个站组成的总体样本的正态性和根据一定原则分区抽样样本的正态性进行研究。

为了对研究对象是否遵从正态分布进行检验,可分别计算偏度系数 $g_1 = \frac{m_3}{m_2^{3/2}}$,峰度系数 $g_2 = \frac{m_4}{m_2^2} - 3$(式中 m_2, m_3, m_4,分别为样本的 2 次、3 次、4 次中心矩,k 阶中心矩表示为 $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$,($k = 2$,

作者简介:康锡言,女,1963 年生,高级工程师,主要从事农业气候区划工作,Email:mhjie64@sina.com

收稿日期:2004 年 2 月 27 日;定稿日期:2004 年 9 月 14 日

3,4),n 为样本数,$\bar{x}$ 为 n 个样本的平均值,x_i 为第 i 个样本),偏度系数的均方差 $s_{g_1}=\sqrt{\frac{6(n-2)}{(n+1)(n+3)}}$,峰度系数的均方差 $s_{g_2}=\sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}$,若显著水平 $\alpha=0.05$,变量 $|g_1|>2|s_{g_1}|$,$|g_2|>2|s_{g_2}|$,则认为变量不遵从正态分布,否则就可以认为变量遵从正态分布[3]。

2 总体样本的正态性分析

首先对 30 个站的年平均气温、经度、纬度、海拔高度求偏度系数和峰度系数(表 1)。

表 1 30 站年平均气温、经度、纬度和偏度、峰度系数

	年平均气温/℃	经度/°E	纬度/°N	海拔高度/m
阜平	12.6	114.18	38.85	281.9
完县	12.4	115.15	38.83	51.0
无极	12.4	114.95	38.18	45.4
平山	12.9	114.20	38.25	131.0
井陉	13.0	114.13	38.03	255.5
新乐	12.4	114.68	38.35	75.1
定县	12.5	115.00	38.52	54.8
安国	12.3	115.33	38.42	29.6
涞源	7.6	114.67	39.35	845.2
赞皇	13.3	114.37	37.65	136.6
临城	13.4	114.38	37.45	113.0
内丘	12.6	114.50	37.28	73.9
隆尧	12.9	114.75	37.35	33.1
巨鹿	13.1	115.03	37.22	29.8
赵县	12.4	114.78	37.75	38.5
宁晋	12.6	114.88	37.62	30.1
栾城	12.7	114.63	37.88	52.9
邢台	13.8	114.50	37.07	77.3
易县	12.0	115.50	39.35	54.6
蔚县	6.9	114.57	39.83	909.5
涉县	12.5	113.67	36.57	470.2
浆水	11.5	113.95	37.17	490.3
峰峰	14.2	114.22	36.42	126.6
武安	12.8	114.15	36.68	233.7
沙河	13.1	114.47	36.87	67.9
磁县	13.1	114.38	36.38	69.7
肥乡	13.1	114.80	36.55	50.2
曲周	13.2	114.95	36.77	39.6
曲阳	12.2	114.68	38.63	104.1
行唐	12.0	114.55	38.45	96.2
偏度系数	-2.72	0.028	0.28	2.28
峰度系数	7.17	-0.104	-0.81	4.29

否定判据 $2s_{g_1}=0.81$,$2s_{g_2}=1.40$,从表中可见,年平均气温、海拔高度的偏度、峰度系数均未通过检验,经度、纬度通过检验,即 30 个站的年平均气温、海拔高度不遵从正态分布,而经度、纬度遵从正态分布。

由于在农业气候区划中,要建立的是 $Y=\sum CX$ 的空间回归模型,Y 为年平均气温(预报量),C 为系数,X 是海拔高度或经度、纬度等地理参数(预报因子)。如果 Y 和 X 为正态变量,当给定山区某点的 X 时,则 Y 的条件分布是正态的[2]。从表 1 可见,30 个站的年平均气温(Y)不遵从正态分布,因此即使经度、纬度(X)遵从正态分布,用回归模型计算得到的年平均气温也是不可靠的。

3 分区样本的正态性分析

由以上分析可知,30 个站总体样本的年平均气温不遵从正态分布,那么要建立空间回归模型,就要从总体样本中按照某种原则抽取样本组成新的序列,重新检验各要素的偏度、峰度系数。考虑到在农业气候区划中建立的是区域较小的回归模型,在抽取样本时站点经纬度的跨度不宜太大,我们把 30 个站按照经纬度跨度不超过 1.5°的标准,把研究区分为 1 区(南部区)、2 区(中南部区)、3 区(中北部区)、4 区(北部区)。经过检验只有 1 区各要素通过正态性检验,1 区站点包括:涉县、浆水、峰峰、武安、沙河、磁县、肥乡、曲周、内丘。为了使其它 3 个区也通过正态检验,分区内的站点必须进行调整,加入相邻区临近本区的站点,然后剔除这组数据中年平均气温偏离均值较远的站点,再进行正态检验,经过反复试验,最终确定:2 区站点包括内丘、临城、栾城、赞皇、赵县、井陉、隆尧,3 区包括行唐、曲阳、新乐、平山、井陉、定县、安国、无极、完县、阜平,4 区包括易县、曲阳、行唐、完县、定县、新乐、平山,能使年平均气温及海拔高度、经纬度通过正态性检验。有些站点被重复分入两个区,建立回归模型时,可以使用这两组数据分别建模,至于两个模型的准确性还需进行检验确定。从表 1 可知,1 区、2 区、3 区、4 区经纬度最大跨度分别为 1.28°E、0.9°N,0.65°E、0.75°N,1.20°E、0.82°N,1.30°E、1.10°N,均未超过 1.5°。表 2 给出了 4 个分区各要素的偏度、峰度系数及其否定判据。

表 2　4 个分区样本的年平均气温、海拔高度、经度、纬度的偏度、峰度系数及否定判据

	1 区		2 区		3 区		4 区	
	偏度系数	峰度系数	偏度系数	峰度系数	偏度系数	峰度系数	偏度系数	峰度系数
否定判据	1.183	1.470	1.225	1.323	1.159	1.509	1.225	1.323
年均气温/℃	-0.211	0.576	0.111	-1.246	0.430	-0.502	0.537	-0.545
海拔高度/m	1.031	-0.656	1.160	0.197	1.108	-0.247	0.493	-1.121
经度/°E	-0.105	-0.740	-0.290	-1.022	0.004	-1.282	0.191	-0.805
纬度/°N	0.592	-0.904	0.121	-1.356	0.125	-0.927	1.086	0.118

从表 2 可以看出,除 2 区纬度的峰度系数不能通过正态检验外,其它均通过检验,因此按照回归分析的要求,可以使用 4 个分区通过正态性检验的数据建立回归模型,但通过正态检验的经度、纬度、海拔高度能否被选择为预报因子,还要进行相关系数显著性检验(表 3)。

表 3　总体样本和各分区样本的年平均气温与海拔高度、经度、纬度的相关系数

	1 区	2 区	3 区	4 区	合并区	总体
经度/°E	0.409	-0.617	-0.568	-0.527	-0.597	-0.083
纬度/°N	-0.653	-0.077	-0.412	-0.596	-0.784	-0.660
海拔高度/m	-0.684	0.493	0.584	0.385	-0.955	-0.841
检验值	0.666	0.754	0.632	0.754	0.666	0.361

注:表中相关系数检验值是信度为 0.05 的临界值。

从表 3 可知,1 区、2 区、3 区、4 区的年平均气温与经度、纬度的相关系数均未通过显著性检验,说明在小区域经度、纬度不是影响年平均气温的主要因素;年平均气温与海拔高度的相关系数只有 1 区通过显著性检验且为负相关,符合山地气温的分布规律,即:在山地,在相同地形条件下,一般都是随着海拔高度升高,空气温度降低[4];其它 3 区年平均气温与海拔高度的相关系数均未通过显著性检验,这一现象与分区所选站点的海拔高度以及站点受地形的影响关系密切。

从表 1 可知,1 区山区站较多且海拔高度差异较大,并且本区受太行山大地形对低丘东缘一带增温效应的影响较小[5],因此 1 区的站点能较好的反映山区气温的特点,可以使用该区数据建立回归模型。2、3、4 区平原站较多且山区站海拔高度较低,并且这些站点大多位于太行山中段,受太行山大地形增温效应影响较大,使低山丘陵气温高于东部低平原[5]。因此要建立适宜于山区的年平均气温的回归模型,必须从这 3 个区中去掉较多的平原站,增加山区站的个数,并且所选站点要具有相同的气候特点,即剔除具有明显局地小气候的站点(例如选站时要避免选城市站,因为受城市的影响,市区气温明显偏高)。经过反复试验,最终从 3 个区中选取的井陉、阜平、平山、赞皇、临城、行唐、涞源、蔚县、浆水组成的空间序列的年平均气温、海拔高度、经度、纬度通过正态性检验,年平均气温与海拔高度、纬度的相关系数通过显著性检验(表 3),因此可以使用该合并区数据建立回归模型。

合并区经度跨度为 0.72° E、纬度跨度较大为 2.66° N,不符合经纬度跨度不超过 1.5°的标准。因此从满足条件的两区站点的性质可以看出,在农业气候区划中建立回归模型时,选取站点应遵从这样的原则:经纬度跨度不超过 1.5°,多选取山区站;或尽量多选山区站,并使同一区的站点具有相同的气候特点,无气温异常站选入,此时可不考虑经纬度跨度的限制。

4　回归模型的残差分析

为了进一步说明进行正态性检验的必要性,把不遵从正态分布的总体样本和遵从正态分布的两个分区的样本,应用线性回归分析方法建立回归模型,并对模型的残差进行分析,以检验模型的拟合效果(图 1)。从表 3 可以看出,总体、1 区、合并区 3 区中,年平均气温与海拔高度相关系数的绝对值最大,为了分析方便,模型中仅选择了海拔高度为预报因子。

由于残差是反映预报量与回归估计值之间的差值大小,因而通常可以利用残差分析来诊断回归模型线性化程度。作残差与回归值的散布图,从图 1a 中可以看出,总体样本的散布点并不是围绕 x 轴密

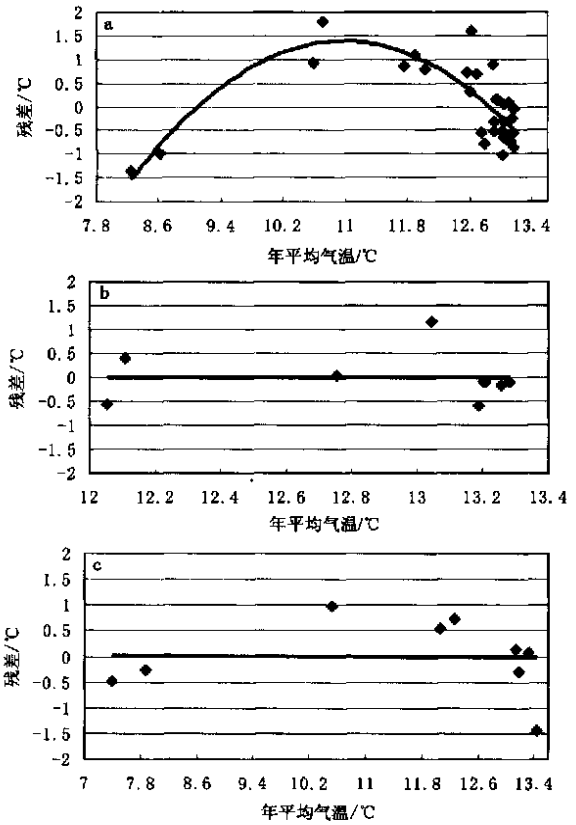

图 1 总体样本和两个分区样本的残差与拟合值的散布图
(a) 总体 ,(b)1 区 ,(c) 合并区

集分布 ,其分布更象是一条曲线 ,说明基于简单线性回归的分析是不正确的 ,因此在变量遵从正态分布假设不成立时 ,即使相关系数和回归方程都通过显著性检验 ,残差仍太大 ,模型不可靠 。从图 1b 、图 1c 可以看出 ,1 区和合并区的散布点比较密集于平行 x 轴的一直线上 ,两图残差绝对值大于 0.5 的分别有 3 个点 ,最大值分别为 1.15 和 1.45 ,说明模型的描述对于大部分数据是正确的 ,回归模型较好 。因此在建立回归模型的过程中 ,相关系数的检验 、回归方程的检验 ,均是在变量遵从正态分布这一条件下进行的 ,对变量进行正态性检验 ,是建立较好回归模型的基础 ,是必须进行的一项工作 。

5 结论

(1) 河北省西部地区 30 个国家基本站的经度 、纬度遵从正态分布 ,年平均气温 、海拔高度不遵从正态分布 。

(2) 根据经纬度跨度不大于 1.5°的标准选取站点 ,能改变年平均气温 、经度 、纬度 、海拔高度的偏度 、峰度系数 ,使其通过正态性检验 ,但有 3 个区的地理参数与年平均气温不相关 。

(3) 在农业气候区划中 ,选取站点应遵从 :经纬度跨度不大于 1.5° ,尽量多选取山区站 ;或既要多选山区站 ,又要使同一区的站点具有相同的气候特点 ,不选入气温异常站两种原则 。

(4) 建立回归模型时 ,首先要检验变量的正态性 ,相关系数检验 、回归方程检验 ,均是在变量遵从正态分布这一前提条件下进行的 ;若前提条件不成立 ,则回归模型残差较大 ,影响预测结果的正确性 。

在农业气候区划中 ,要建立气温 、降水 、太阳辐射等多要素的空间回归模型 。假定变量遵从正态分布不进行检验 ,仅对相关系数 、回归模型进行显著性检验 ,得到的回归直线误差较大 ,不能较准确推算山区某点的气象要素 ,因此对变量进行正态性检验是必须进行的一项工作 。

参考文献

1 王建源 , 冯晓云 , 薛德强 , 等 . GIS 在泰安市板栗农业气候区划中的应用 . 中国农业资源与区划 , 2003 , (5) : 47
2 Weisberg S . 应用线性回归(第二版) . 北京 : 中国统计出版社 , 1998 . 75 - 76
3 黄嘉佑 .气象统计分析与预报方法(第二版) . 北京 : 气象出版社 , 2000 .18
4 傅抱璞 . 山地气候 . 北京 : 科学出版社 , 1983 . 114
5 程树林 , 郭迎春 , 郭康 . 太行山燕山气候考察研究 . 北京 : 气象出版社 , 1993 . 35 - 36

(下转第 192 页)

(上接第 188 页)

Normality of Annual Mean Temperature in Climate Regionalization in Mountainous Regions

Kang Xiyan Gao Jianhua

(Hebei Institute of Meteorology Science, Shijiazhuang 050021)

Abstract: Annual mean temperatures of 30 years from 30 different national basic weather stations in the western part of Hebei Province were used, and through analyzing the skewness coefficient and kurtosis coefficient of the population samples, the normality of annual mean temperature and geographic parameters (altitude, longitude and latitude) was studied. Using the ensemble data, the annual mean temperature and altitude did not follow the normal distribution, but the longitude and the latitude follow the normal distribution significantly. According to the principles of station selection—with the span of longitude and latitude being less than 1.5° and using as many stations over mountainous areas as possible; using a group sample data with the same climate characteristics; and also using the mountainous areas stations at the same time, the station grouping was conducted, and hereby the annual mean temperature and geographic parameters in the same sample group follow the normal distribution.

Key words: normal distribution, skewness coefficient, kurtosis coefficient, climate regionalization