基于主成分聚类分析的美国黄松引种区划方法

王荣繁，唐德瑞

（西北农林科技大学 林学院，陕西 杨凌 712100）

摘　要：对我国42个美国黄松引种地的年均温、最热月均温、最冷月均温、极端最高温、极端最低温、≥10℃积温、无霜期、年均降雨量、年蒸发量、平均风速、年日照时数、平均海拔12个气候指标进行了主成分分析和聚类分析，划分为最适宜区、适宜区、次适宜区和不适宜区4种类型，并结合美国黄松的生长状况对2种方法区划的结果进行对比。结果表明：主成分-聚类分析法既可以对多变量数据进行合理地分类，又能对各类优劣程度做出综合评价，能充分反映适宜美国黄松生长的气候情况，与美国黄松实际生长情况比较后，验证了该方法是切实可行的。

关键词：美国黄松；引种；主成分分析；聚类分析

中图分类号：S 722.7　**文献标识码**：A　**文章编号**：1001－0009(2012)18－0092－04

美国黄松(*Pinus ponderosa*)是北美西部分布最广泛的树种之一。其树皮粗厚，树干高大通直，木材坚硬，材质好，是优良建筑用材和防火树种，也是干旱地区造林的先锋树种[1]。我国从20世纪30年代就开始了美国黄松的引种工作[2]，目前我国北部辽宁、吉林、内蒙古、北京等地，西北部黄土高原地区以及少数南方地区浙江、上海、长沙等均对美国黄松进行了引种，其中西北地区为引种的主要区域[3-9]。由于引种地域广大，各引种点间气候条件差异悬殊，生长状况差异显著，为了合理规划美国黄松在我国引种区的布局，避免盲目引种，对美国黄松引种地区的气候区划进行研究具有十分重要的意义。

1　材料与方法

1.1　资料的收集

该研究通过查阅中国气象站公布的气候资料，以林木生存主要影响气象因素的温度、水分、光照条件为依据[9]，选取我国42个黄松引种地(图1)的年均温(℃)、最热月均温(℃)、最冷月均温(℃)、极端最高温(℃)、极端最低温(℃)、≥10℃积温(℃)、无霜期(d)、年均降雨量(mm)、年蒸发量(mm)、平均风速(m/s)、年日照时数(h)、平均海拔(m)12个关键气候因子作为美国黄松气候生态范围研究的比较因素(表1)。由于美国黄松原产地的分布十分广泛，由北向南从加拿大南部穿过美国一直到达墨西哥[10-11]，因此该研究中关于原产地的数据为各地气候的平均值。

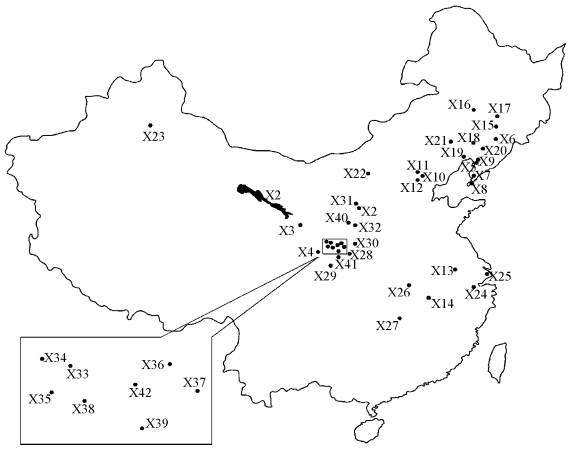

图1　42个引种地的分布情况

Fig.1　Distributions of 42 introduced areas of China

调查收集引种种植区美国黄松的年均树高生长量(m)和年均胸径生长量(cm)2个关键生长因子作为生长评价的指标(表2)。

1.2　统计分析方法

主成分-聚类分析法是利用SPSS 16.0将美国黄松引种地的气候因子和生长指标中的大部分信息通过少数几个综合指标反映出来，再取若干主成分对样品进行聚类分析，结合第一主成分排序对样品进行分类排名，由此得到一种新的综合评价方法。具体步骤[12]如下。

设有 n 个样本，每个样本有 p 项指标(变量) x_1，x_2，…，x_p。

(1)先对数据进行标准化处理，以使每一个变量的平

第一作者简介：王荣繁(1987-)，女，陕西商洛人，硕士，研究方向为森林培育。E-mail：taotie19871214@163.com.

责任作者：唐德瑞(1961-)，男，博士，教授，博士生导师，现主要从事森林培育及生态学的教学和研究工作。E-mail：taotie19871214@163.com.

基金项目：国家林业局“948”资助项目(98-4-05)。

收稿日期：2012－05－23

均值为0,方差为1。X'的计算公式为 $X'=(x_{ip}-\overline{x_p})/s_p$。

其中,$\overline{x_p}$和 s_p 分别为第 j 个变量的平均值和标准差。

(2)计算指标的相关矩阵R,求R的p个特征值记为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0$,相应的正交化特征向量 $v_i=(v_{i1},v_{i2},\cdots,vi_p)$,$i=1,2,\cdots,p$。

(3)设方差贡献率 $\alpha_i=\lambda_i/\sum_{i=1}^{p}\lambda_i$,当累计方差贡献率 $\sum_{i=1}^{p}\alpha_i$ 达到一定的数值(一般≥85%)时,取q个主成分 $y_i=v_{i1}x_1+v_{i2}x_2+\cdots+v_{ip}x_p(i=1,2,\cdots,q)$,进而得到综合评价函数:$Y=(\alpha_1\overline{y}_1+\alpha_2\overline{y}_2+\cdots+\alpha_q\overline{y}_q)/\sum_{i=1}^{p}\alpha_i$。

(4)对选定的新数据阵$(Y_1,Y_2,\cdots,Y_q)$进行系统聚类分析。

表1　42个引种区和原产地的主要气候要素值

Table 1　Values of the main climatic elements of 42 introduced areas of China

引种地	年均温/℃	最热月均温/℃	最冷月均温/℃	极端最高温/℃	极端最低温/℃	≥10℃的积温/℃	无霜期/d	年降雨量/mm	年日照时数/h	年蒸发量/mm	年均风速/m·s^{-1}	平均海拔/m
X1	7.3	29.1	−8.0	38.6	−33.3	3 014.0	171	130.5	3 109.7	2 420.7	2.73	1 500
X2	8.1	28.8	−7.7	38.6	−32.7	3 217.6	161	358.0	2 925.7	1 895.7	2.30	939
X3	9.1	28.4	−4.6	39.1	−21.7	3 242.0	168	368.0	2 607.6	1 437.7	1.00	1 580
X4	10.4	28.4	−4.6	39.0	−21.7	3 242.1	167	368.1	2 608.1	1 437.0	0.80	1 250
X5	9.0	24.4	−7.6	36.6	−28.5	3 516.0	169	687.0	2 819.2	1 609.0	2.70	20
…	…	…	…	…	…	…	…	…	…	…	…	…
X38	12.8	30.3	−2.9	41.7	−16.1	3 462.0	197	350.0	1 928.1	1 383.2	1.23	950
X39	13.1	30.3	−2.9	41.7	−16.0	4 053.7	209	352.0	1 928.8	1 770.3	1.24	1 050
X40	9.8	29.1	−4.4	39.7	−25.1	3 270.8	180	362.0	2 449.7	1 556.0	1.90	1 300
X41	12.9	30.7	0.4	42.0	−19.4	4 169.2	225	397.0	2 163.0	1 505.0	2.10	1 600
X42	13.2	30.3	−2.8	41.7	−19.1	4 108.9	210	351.0	1 928.9	1 337.0	1.30	1 410
X43	8.3	19.7	−2.8	41.6	−33.6	2 053.0	145	437.5	2 696.44	1 500.0	3.20	482.9

注:X43为美国原产地气候要素的平均值。

Note:X43 is the average climate elements of origins in the United States.

表2　42个引种区的主要生长要素值

Table 2　Values of the main growth elements of 42 introduced areas of China

引种地	年均树高生长量/m	年均胸径生长量/cm	引种地	年均树高生长量/m	年均胸径生长量/cm	引种地	年均树高生长量/m	年均胸径生长量/cm
X1	0.12	0.39	X15	0.10	0.56	X29	0.16	0.34
X2	0.22	0.53	X16	0.10	0.23	X30	0.21	0.36
X3	0.11	0.38	X17	0.08	0.22	X31	0.10	0.31
X4	0.12	0.39	X18	0.08	0.23	X32	0.18	0.38
X5	0.30	0.54	X19	0.08	0.22	X33	0.45	0.52
X6	0.22	0.31	X20	0.08	0.21	X34	0.22	0.36
X7	0.23	0.31	X21	0.11	0.52	X35	0.19	0.36
X8	0.23	0.34	X22	0.16	0.46	X36	0.22	0.53
X9	0.23	0.37	X23	0.13	0.25	X37	0.21	0.36
X10	0.41	0.67	X24	0.02	0.11	X38	0.20	0.53
X11	0.09	0.27	X25	0.03	0.12	X39	0.13	0.28
X12	0.34	0.45	X26	0.05	0.13	X40	0.24	0.30
X13	0.07	0.18	X27	0.06	0.14	X41	0.24	0.45
X14	0.07	0.12	X28	0.23	0.60	X42	0.16	0.34

表3　各指标的Pearson相关系数矩阵

Table 3　Pearson correlation coefficient matrix of each index

	年均温/℃	最热月均温/℃	最冷月均温/℃	年降雨量/mm	极端最高温/℃	极端最低温/℃	无霜期/d	≥10℃的积温/℃	年日照时数/h	年蒸发量/mm	年均风速/m·s^{-1}	平均海拔/m
年均温/℃	1.000											
最热月均温/℃	0.533	1.000										
最冷月均温/℃	0.919	0.406	1.000									
年降雨/mm	0.396	−0.257	0.429	1.000								
极端最高温/℃	0.447	0.601	0.386	−0.290	1.000							
极端最低温/℃	0.871	0.562	0.810	0.320	0.335	1.000						
无霜期/d	0.868	0.403	0.871	0.410	0.351	0.823	1.000					
≥10℃积温/℃	0.846	0.388	0.852	0.479	0.442	0.748	0.779	1.000				
年日照时数/h	−0.784	−0.385	−0.706	−0.378	−0.343	−0.771	−0.764	−0.682	1.000			
年蒸发量/mm	−0.452	0.011	−0.532	−0.448	0.074	−0.470	−0.506	−0.419	0.656	1.000		
年均风速/m·s^{-1}	0.231	−0.247	0.268	0.710	−0.305	0.152	0.271	0.377	−0.050	−0.020	1.000	
平均海拔/m	−0.045	0.330	0.049	−0.543	0.252	0.038	0.058	−0.219	−0.006	−0.066	−0.607	1.000

2　结果与分析

2.1　相关性分析

为了检验主成分分析的可行性，首先对各气象因子进行相关性分析。由表 3 可知，年均温度、最冷月均温、极端最低温、无霜期与≥10℃积温这几个气候因子之间的相关系数很高，其中年均温度与最冷月均温之间的相关系数高达 0.919，而最冷月均温与极端最低温，极端最低温与无霜期，无霜期与≥10℃积温之间的相关系数分别为 0.871、0.868 和 0.846，具有明显的相关性，因此有必要对这些数据进行主成分分析。

2.2　主成分-聚类分析

用主成分分析法处理数据后，由表 4 可知，前 3 个主成分方差(λ_i)大于 1，且累积贡献率达 83.189%，故前 3 个主成分是从提取出对解释原有变量贡献最大的主成分，其中第 1 主成分贡献率达 50.475%，根据表 5 中主成分各载荷因子所占权重得出，第 1 主成分主要综合了年均温、最冷月均温、极低温度、≥10℃积温、无霜期、年日照时数 6 个气候因子，其特征向量所凝聚的气象信息主要是低温因素，故称第 1 主成分为低温因子。第 2 主成分贡献率为 22.737%，其气候因素受湿度影响较大，主要集中了最热月均温、极端最高温度、年降雨量、年均风速、平均海拔 5 个气候因子，可命名为高温因子。第 3 主成分贡献率为 9.977%，特征向量仅由年蒸发量来反映，故称第 3 主成分为湿度因子。

表 4　主成分的特征值、方差贡献率和累积贡献率

Table 4　Eigen values, contribution rate and accumulative contribution rate of principal component variance

主成分	提取平方和载入		
	特征值	贡献率/ %	累积贡献率 /%
1	6.057	50.475	50.475
2	2.728	22.737	73.212
3	1.197	9.977	83.189

表 5　主成分因子载荷矩阵

Table 5　Factor load matrix of principal components

气候因子	第 1 主成分	第 2 主成分	第 3 主成分
年均温度/℃	0.958	0.059	0.102
最热月均温/℃	0.492	0.650	0.301
最冷月均温/℃	0.935	0.010	−0.012
年降雨量/mm	0.480	−0.790	−0.072
极端最高温/℃	0.420	0.643	0.399
极端最低温/℃	0.903	0.112	−0.012
无霜期/d	0.920	0.013	−0.043
≥10℃积温/℃	0.899	−0.107	0.225
年日照时数/h	−0.857	−0.060	0.250
年蒸发量/mm	−0.581	0.167	0.722
年均风速/m·s^{-1}	0.271	−0.798	0.353
年均海拔/m	−0.043	0.756	−0.414

最后根据各个地区气候因子以及该地区美国黄松生长状况的得分，分别与对应的方差贡献率相乘，计算出气候因子的综合得分(表 6)，从而给予各引种地的气候条件和美国黄松的生长质量状况以定量化描述。气候因子综合得分越大，表明气候条件越适合美国黄松的生长；生长质量综合得分越高，表明美国黄松的生长效果越好。

表 6　美国黄松气候因子综合指标

Table 6　Comprehensive quality index of climate factors of *Pinus ponderosa*

引种地编号	生长因子		
	第 1 主成分	第 2 主成分	综合指标值
*X*1	0.25689	4.965	5.22189
*X*2	0.37791	17.874	18.25191
*X*3	0.24680	4.965	5.21180
*X*4	0.25689	4.965	5.22189
*X*5	0.42342	42.699	43.12242
…	…	…	…
*X*38	0.28734	9.930	10.21734
*X*39	0.36779	6.951	7.31879
*X*40	0.20662	6.951	7.15762
*X*41	0.27234	8.937	9.20934
*X*42	0.34779	10.923	11.27079

最后采用系统聚类法中的最近邻距离法，在 SPSS 16.0 中对上述前 3 个主成分得分数据矩阵做聚类分析，42 个美国黄松的引种地被分为 4 类：最适宜区{*X*1、*X*2、*X*3、*X*17、*X*18、*X*19、*X*22、*X*31}；次适宜区{*X*6、*X*10、*X*14、*X*20、*X*28、*X*33、*X*34、*X*35、*X*38、*X*41、*X*42}；不适宜区{*X*13、*X*24、*X*25、*X*26、*X*27、*X*29}；其它地区都属于适宜区。

2.3　结果的对比

为了增强说服力，根据美国黄松的实际生长情况，利用主成分分析法进行分类，并适当照顾区域的连续性，将美国黄松引种地划分为以下 4 种类型(图 2)：生长优良的地区为河西走廊、兰州徐家山、甘肃陇南、辽宁等地、北京八达岭附近的石质山地；生长良好的地区为西北黄土高原中部，华北地区；可生长地区为吉林、辽宁、沈阳、乌鲁木齐、陕西各地；美国黄松生长较差的地区为浙江富阳、上海、武昌、湖南长沙、江西庐山、江苏南京。

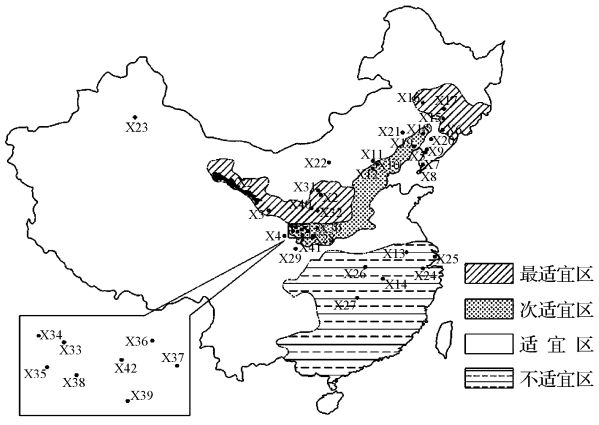

图 2　美国黄松在我国引种地区划示意图

Fig. 2　Division for suitable climatic and ecological regions of continental climatic type of *Pinus ponderosa* in China

主成分方法的分类结果与实际生长情况相比,虽然二者的不适宜区划分结果基本相符,但主成分分析只把X1、X19 2个引种点归类为最适宜区,而且其余的大部分地区属于适宜区,故利用主成分分析法划分的结果存在较大误差。

研究发现,当主成分分析和聚类分析2种方法结合起来时,划分的结果为:最适宜区{X1、X2、X3、X17、X18、X19、X22、X31};次适宜区{X6、X10、X14、X20、X28、X33、X34、X35、X38、X41、X42};不适宜区{X13、X24、X25、X26、X27、X29};其它地区都属于适宜区,与依据美国黄松实际生长状况划分的结果十分接近。

表7　　2种分析方法分类的结果

Table 7　Classification results of three divisional methods

	最适宜区	适宜区	次适宜区	不适宜区
实际生长情况	1,3,4,6,7,8,9,11	2,5,10,12,14,17,19,21,22,26,28,3234,37,39,40,42	15,16,18,20,23,30,31,35,36,38,41	13,24,25,27,29,33
主成分分析	1,19	2,3,4,5,7,8,9,10,11,12,15,16,17,18,21,22,23,28,30,31,32,33,34,35,36,37,38,39,40,41,42	6,14,20	13,24,25,26,27,29
主成分-聚类分析	1,2,3,17,18,19,22,31	4,5,7,8,9,11,12,15,16,20,21,23,30,32,36,37,39,40	6,10,14,20,28,33,34,35,38,41,42	13,24,25,26,27,29

注:表中数字为引种地的编号。

Note:Figures in the table is the numbers of introduced areas.

3　结论与讨论

该文以我国42个美国黄松引种地的气象数据为依据,分别用主成分分析法和主成分-聚类分析方法对我国引种美国黄松的区域进行划分,并与依据生长表现划分的气候适生区进行比较。主成分分析法能将影响美国黄松引种的气候因素或指标简化为少数几个含义可解释的综合指标,起到抓住问题实质的作用,并可通过主成分得分的高低来量化各因素的优劣性,它以主成分的贡献率作为权重,权重的获得比较简便客观,分类评价指标构造简单、易操作。聚类分析则能兼顾对象多因素的联系和主导作用,可按它们的亲疏差异程度逐步分组归类,更能客观地反映变量或区域之间的内在组合关系。将主成分分析法以及聚类分析二者结合对美国黄松的适宜引种区进行分类比较,得出在对适宜引种的地区提出可行性预测时,更能真实的反映数据的本质特征,其结果也与美国黄松的实际生长情况最吻合,这将为优选适宜美国黄松生长地区开辟一条新的途径,具有广阔的应用前景。

参考文献

[1] Jim W O, Russell A R. Silvies of Forest Trees of the United States [M]. 1998:417-430.

[2] 张廷福. 辽宁森林[M]. 北京:中国林业出版社,1990:252-254.

[3] 张立功,王喜武,张仁慈,等. 黄松引种研究[J]. 东北林业大学学报,1997,25(2):9-12.

[4] 罗伟祥,宋西德,侯琳,等. 黄土高原美国黄松引种生长调查研究[J]. 陕西林业科技,1998(1):1-8,12.

[5] 周永学,樊军锋,高建社,等. 美国黄松在陕西黄土丘陵山地引种效果分析[J]. 西北农林科技大学学报(自然科学版),2005,33(4):83-86.

[6] 周永学,樊军锋,杨培华,等. 陕西陇县引种美国黄松生长调查[J]. 西北林学院学报,2005,20(3):74-77.

[7] 陈斌,葛宏元. 河西走廊引种栽培美国黄松试验[J]. 林业实用技术,2006(8):42.

[8] 白红霞,陈延,曹锋. 黑龙潭树木园引种美国黄松生长调查[J]. 陕西林业科技,2008(1):80-81,91.

[9] 吴中伦. 国外树种引种概论[M]. 北京:科学出版社,1983:135-136.

[10] Scott R A, Wallace W C. Forest ecosystems of an Arizona *Pinus ponderosa* landscape: multifactor classification and implications for ecological restoration [J]. Journal of Biogeography (J Biogeogr),2006(33):1368-1383.

[11] Jodi R N, Stephen T J, Julio L B. Classification tree and minimum-volume ellipsoid analyses of the distribution of ponderosa pine in the western USA [J]. Journal of Biogeography (J Biogeogr),2006(33):342-360.

[12] 何晓敏. 现代统计分析方法与应用[M]. 北京:中国人民大学出版社,2005:139-143.

Optimization of Divisional Methods Based on Principal Composition Cluster Analysis on Suitable Ecological Region for Introduced *Pinus ponderosa*

WANG Rong-fan, TANG De-rui

(College of Forestry, Northwest Agriculture and Forestry University, Yangling, Shaanxi 712100)

Abstract: Principal component analysis and cluster analysis were used to study 12 climate indicators of *Pinus ponderosa* in 42 places in China, such as annual average temperature, temperature in the hottest and coldest month, extreme maximum and minimum temperature, ≥10℃ accumulated temperature, frost-free period, average annual rainfall and evaporation, the average wind speed, the average wind speed and annual sunshine hours and average altitude. They were divided into 4 types, the most suitable area, suitable area, sub-suitable area and unsuitable area, and comparisons were made among two analytic methods combined with the actual growth of Pinus ponderosa. The results showed that the principal component-cluster analysis not only can classify multivariate data, but also is able to make a comprehensive evaluation; it correctly reflects the most suitable areas for *Pinus ponderosa*. After testing by the actual growth of *Pinus ponderosa*, we confirm the method was practicable.

Key words: *Pinus ponderosa*; introduction; principal component analysis; cluster analysis