

用主组元聚类法判别广东 杉木立地类型

何昭珩 刘有美 李侯才

(林学系)

(农机系)

提 要

本文用一种简便的主组元聚类法对广东杉木立地进行分类。在选取分类指标的方法上作了改进,选取8个对杉木生长影响较大的环境因子为分类指标,计算了这些因子的相关矩阵前4个较大特征值的主组元,选取贡献率较大而又与立地指数显著相关的第2、3主组元为分类的综合指标,作平面聚类图,将40个样本分为6类,从中辨认出各类生产力,结果表明:

第1、2类立地指数较大,适宜种杉。第3、4类生产力中等,可以种杉。第5类严重缺钾,需增加土壤中的含钾量,以提高生产力。第6类因部位不适合,不宜种杉。

本研究结果,不仅为广东杉木生产估计经济效益提供了科学依据,同时也为分类工作者提供了一种方法。

一、主组元的数学模型

为了消除单位的影响,按下式把表示环境的变量 x_{ij} 标准化:

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{S_j} \quad \text{简记 } x_{ij}^* \text{ 为 } x_{ij}, \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, m)$$

n 是样本单元数, m 是环境因子个数。

x_{ij} 是第 i 个样本的第 j 个指标(因子)的值。

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \quad S_j = \sqrt{\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{(n-1)}}$$

表示环境因子的标准化变量 x_{ij}^* 的协方差即相应的原变量 x_{ij} 的简相关系数。记 r_{ij} 为第 j 个因子与第 i 个因子的简相关系数,这些相关系数构成的相关矩阵记为 R ,即

$$R = (r_{ji})_{m \times m}$$

由于 R 的特征值 λ_t ($t = 1, 2, \dots, m$) 为非负实数,故可设 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$

又由于 R 的特征值可取 m 个实的非零正交向量,把这 m 个正交向量正规化即得 m 个正交正规化特征向量 $\zeta_1, \zeta_2, \dots, \zeta_m$

其中第七个特征向量 ζ_t ($t = 1, 2, \dots$) 的第 j 个分量(支量)记为 ζ_{jt} ($j = 1, 2, \dots, m$),并令 $P = (\zeta_{jt})_{m \times m}$, P 是一个 m 阶方阵,则有:

$$P'RP = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_m \end{pmatrix} \dots\dots\dots (1)$$

记这个对角阵为 Λ 。因 R 与对角阵 Λ 相似,故有:

$$\lambda_1 + \lambda_2 + \cdots + \lambda_m = r_1 + r_2 + \cdots + r_m = m$$

主组元就是以特征值 λ_i 的特征向量的分量为权系数的各环境因子的线性组合。

如令 λ_i 的主组元为 y_i ,则

$$y_i = \sum_{j=1}^m x_j \cdot \xi_{ij} \cdots \cdots \cdots (2)$$

主组元是各因子的综合指标,权系数的绝对值的大小和权系数的正负号,分别反映相应环境因子对主组元的作用大小和作用方向。

若令各样本单元的主组元值 y_{it} 为元素的矩阵为 $Y = (y_{it})_{n \times m}$,则:

$$\frac{1}{n-1} Y'Y = \Lambda \cdots \cdots \cdots (3)$$

(3)式表明:主组元的协方差矩阵等于对角阵 Λ ,故知任两主组元是不相关的,因而,用主组元代替原来因子,可以消除由于因子间的相关而产生的信息重叠,所以,往往用一个或少许几个主组元就能反映原来因子的较多信息。

由(3)式还可看出:主组元 y_i 的方差即特征值 λ_i ,故 y_i 对总方差的贡献率为 $\frac{\lambda_i}{m}$ 。

前 K 个主组元对总方差的累计贡献为

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_K}{K}$$

当前 K 个主组元的累计贡献率达到70%以上时,就认为这 K 个主组元能把原来因提供的全部信息的70%以上反映出来。

在聚类分析中,通常是取前面一个或若干个贡献率较大的主组元作为分类的综合指标。如当前 K 个贡献率最大的主组元的累计贡献率达到一定水平时,则取这些主组元作为聚类分析的依据。

二、主组元的电算结果与分析

(一) 相关矩阵 ($R = [r_{ji}]_{8 \times 8}$)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
	(持水量)	(容量)	(全N)	(全K)	(pH)	(粉粒)	(部位)	(坡度)
$R =$	1	-.7651	.3180	-.2839	.1719	.2224	.1614	-.00862
	-.7651	1	-.3745	.05425	-.2557	-.2524	-.00922	.3348
	.3180	-.3745	1	.1485	.2414	.01892	.08746	.1373
	-.2839	.05425	.1485	1	-.0156	-.0582	.04183	-.02828
	.1719	-.2557	.2414	-.01560	1	.2733	-.1010	-.3457
	.2224	-.2524	.01892	-.05820	.2733	1	-.3943	-.1617
	.1614	-.009223	.08746	.04183	-.1010	-.3943	1	.03417
	-.008621	.3348	.1373	-.02828	-.3457	-.1617	.03417	1

(二) 特征值与贡献率

求出R的特征值并把它们由大到小排列, 算出前4个最大特征值的贡献率及其累计贡献率见表1:

(三) 特征向量与因子组合

求出各特征值 λ_i 的特征向量, 并把它们单位化。下面列出了前4个主组元的正交规范化特征向量及其主要因子组合(表2)。

表1 特征值与贡献率

特征值序号	特征值	贡献率 (%)	累计贡献率 (%)
1	2.34726	29.3408	29.3408
2	1.52133	19.0166	48.3573
3	1.19851	14.9814	63.3387
4	1.08088	13.511	76.8497

表2 前四个主组元的特征向量及主要因子组合

NO	主组元 因子	1	2	3	4
1	x_1 持水量	0.51774	0.31430	-0.29505	0.0073933
2	x_2 容重	-0.5736	-0.14401	-0.00062067	0.075001
3	x_3 全N	0.31096	0.35062	0.342676	0.46193
4	x_4 全K	-0.11208	0.0023259	0.76243	0.21084
5	x_5 pH	0.36232	-0.25686	0.31667	-0.10061
6	x_6 粉粒	0.31771	-0.47201	-0.13908	0.30167
7	x_7 部位	-0.043652	0.58852	0.13978	-0.45334
8	x_8 坡度	-0.24404	0.34987	-0.27397	0.65562
主要因子组合	正向因子	持水量 pH N 粉粒	部位 全N 坡度、持水量	全K 全N pH	全N、坡度
	逆向因子	容重	粉粒		部位

上表中第t主组元 ($t = 1, 2, 3, 4$) 的特征向量是相关阵R的特征值 λ_i 的特征向量, 它的各个分量反映各因子对主组元的作用大小和方向, 从表中提供的信息可对各个主组元的性质作如下分析:

第1主组元: 持水量有较大的正向作用, 容重有较大的逆向作用, 所以第1主组元主要反映土壤的疏松状况与持水能力。

第2主组元: 部位、含氮量、坡度、持水量有较大的正向作用, 而粉粒则有较大的负向作用, 故第2主组主要反映地形条件和土壤保持水、肥的能力。

第3主组元: 钾、氮、pH值均有较大的正向作用, 故第3主组可看成是土壤养分和酸碱度的测度。

第4主组元: 坡度、氮、有较大的正向作用, 部位有较大的负向作用, 故第4主组元是地形和含氮量的测度。

(四) 主组元方程

$$y_1 = 0.130306x_1 - 3.78246x_2 + 5.19036x_3 - 0.0674081x_4 + 1.42598x_5$$

$$+ 0.0329418x_6 - 0.0385621x_7 - 0.525838x_8 - 6.37285$$

$$y_2 = 0.0791038x_1 - 0.949649x_2 + 5.85238x_3 + 0.00139885x_4 - 1.01091x_5$$

$$- 0.0489405x_6 + 0.5199x_7 + 0.753874x_8 + 0.415195$$

$$y_3 = -0.0742593x_1 - 0.00409288x_2 + 5.71983x_3 + 0.458535x_4 + 1.24632x_5$$

$$- 0.0144209x_6 + 0.123486x_7 - 0.590329x_8 - 4.34765$$

$$y_4 = 0.00186075x_1 + 0.494577x_2 + 7.71032x_3 + 0.126799x_4 - 0.395964x_5$$

$$+ 0.0312794x_6 - 0.40048x_7 + 1.41269x_8 - 4.86732$$

主组元方程既是确定各样本点坐标位置的依据, 也为求算未知样点的类属提供了可能。

(五) 由上述方程求得40个样品的4个主组元的值(表8)

三、聚类分析方法与杉木立地类型分类

(一) 主组元与目的指标的相关关系

为了探讨立地指数Z与各个主组元的关系, 求出了前4个主组元与立地指数的相关系数如下:

$$r_{1z} = 0.1965$$

$$r_{2z} = 0.3132^*$$

$$r_{3z} = 0.3660^*$$

$$r_{4z} = 0.4259^{**}$$

检验假设总体相关系数 $\beta = 0$ 时, 其5%水准与1%水准的临界值分别为 $r_{0.05} = 0.3124$, $r_{0.01} = 0.398$ 。

r_{ij} 表示第j个主组元与立地指数Z的简相关系数。右上角加*号者, 表示在5%水平显著; 加**号者, 表示在1%水平显著。

上述结果表明: 第1主组元与立地指数的相关关系不显著。可见第1主组元的贡献率虽然最大, 但它所贡献的许多信息却不是这次分析的目的最需要的。除第1主组元外, 其他三个主组元与杉木立地指数的相关达到显著或极显著, 可以考虑以这三个主组元作为分类的依据。但由于三维空间作图不如平面图直观, 所以试取其中贡献率较大的第2、3主组元作为分类的综合指标。

(二) 聚类分析图

分别以第2、3主组元为横纵坐标, 将表8中40对数据描绘在图1中, 图中●点叫样

表 8 40 个 样 品 的 主 组 元 值

NO	y_1	y_2	y_3	y_4
1	2.69516	-1.02910	1.53070	0.534362
2	-0.282903	0.694258	1.67384	1.55768
3	-0.178224	-0.121442	1.50469	-0.608972
4	0.791212	-0.956563	-0.189840	-0.278662
5	2.87285	-0.321757	-0.102760	0.548887
6	0.871291	-0.821621	0.237092	-0.382122
7	0.749761	-1.21869	0.534741	0.994229
8	0.221681	-2.08141	0.243208	0.0281995
9	0.971067	-0.213198	-1.07414	2.28161
10	2.47434	0.44180	0.0304404	0.260516
11	2.91216	-0.867932	0.331523	2.39584×10^{-3}
12	-0.519873	0.216467	-1.07432	1.19414
13	0.0609164	1.11301	0.651702	2.09083
14	0.42410	2.89485	0.566082	0.980409
15	-0.133033	-0.134624	-0.497828	-0.779577
16	-1.44534	-0.349636	0.564116	-0.532029
17	1.22455	0.269524	-0.0508396	-0.0554834
18	-1.64428	0.336397	-0.0713427	1.4300
19	-0.176512	1.99110	2.09689	-1.32070
20	-5.79372×10^{-3}	0.386552	2.22640	-0.192633
21	-1.93096	-0.933431	0.507718	-0.886332
22	-1.28788	1.09652	-1.51562	-0.113698
23	-0.823624	0.175389	-0.342095	-1.37569
24	0.724403	-1.61191	-2.24063	-1.08509
25	0.543661	-0.34795	-0.606412	-1.64570
26	-2.14555	1.52667	-0.646473	-0.320167
27	-0.603425	-1.44024	-1.10322	-1.22904
28	-0.292245	0.440527	-0.0373921	-1.07256
29	2.93194	2.42993	-0.200854	-0.416937
30	-2.96497	-0.236394	-1.69822	0.932066
31	-0.852615	2.06911	-2.03426	1.02290
32	-0.680862	1.63962	-0.819267	-0.248828
33	2.79174	0.759685	-1.21047	-1.14889
34	-0.786651	0.905717	-0.374635	-1.48634
35	-1.98566	0.380662	1.60847	-1.17450
36	-2.44653	-1.20691	1.90348	0.153351
37	0.365933	-2.34238	-0.672972	0.513265
38	-0.464437	-0.953183	-0.0353634	-0.199896
39	-1.44510	-1.50344	0.0988894	2.01943
40	-0.530293	-1.07597	0.28896	9.55564×10^{-3}

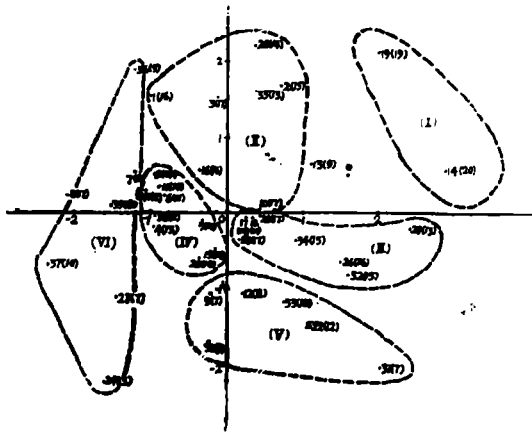


图1 主组元聚类

图1中的横坐标为第2主组元,纵坐标为第3主组元,样品点越向右,则表示地形条件、氮素、水分越有利;样品点越向上,则表示钾、氮、养分和pH值越有利。又由于第2主组元不仅反映地形因子的信息,还反映含氮量和持水量的信息,故样品点在横坐标方向上移动比在纵坐标方向上移动更为关键。图1中每条围线内的样品点可看成为一类,因为它们的主要因子水平和生产力比较一致。故全省杉木立地类型可划分为如下六类:

第Ⅰ类:谷地多湿肥沃酸性型。包含14号、19号两个样品。在坐标系的最右边中上部,为最有利于杉木生长的环境,属高肥力型,立地指数平均值 $\bar{y}_1 = 19.5$ 。

第Ⅱ类:坡地肥沃酸性型。样品序号依次为:1、2、3、10、16、20、35,位于坐标系中间上部。地形条件中等,养分、pH值最为有利,故其生产力也较高,立地指数平均值 $\bar{y}_2 = 15.4$ 。

第Ⅲ类:坡地潮湿酸性型。样品序号为17、18、23、26、28、29、32、34,位于坐标系中间偏右,各种条件中等稍偏高生产力,中等偏高,立地指数平均值 $\bar{y}_3 = 14.5$ 。

第Ⅳ类:陡坡中肥酸性型。样品序号为4、5、6、11、15、21、25、38、40,位于坐标系中间偏左,条件中等,生产力中等,立地指数平均值 $\bar{y}_4 = 12.7$ 。

第Ⅴ类:坡地贫钾酸性型。样品序号为9、12、22、30、31、33,位于坐标系下部右方,虽然地形条件尚好,但养分缺乏,对杉木生长不利,生产力低,立地指数平均值仅8.8。不宜种杉。

第Ⅵ类:高亢干旱型。样品序号为7、24、27、36、37、39,位于坐标系最左边,主要是地形部位不利、强光照、持水能力低,故即使其他条件并不太差,但杉木生长仍很差,生产力最低,立地指数平均值为8.8。也不宜种杉。

综上所述,Ⅰ、Ⅱ类立地为高肥力宜杉型,Ⅴ、Ⅵ类立地为低肥力不宜杉型,Ⅲ、Ⅳ类立地为中肥力过度类型。各类立地的环境条件和土壤指标列入表4(过度类未列入)。

品点,它表示样品的位值。如14号样品第2、第3主组元值依次为2.89485,0.566082,它的坐标位置在第一象限最右边而与横坐标轴较为接近(图1)。图中圆点旁的数字14表示样本序号,其旁括号内的数字20表示该样品的立地指数为20。这是生长最好的样品。40个样品点大致聚在6条围线内。其中仅两个点不在围线内,这可能是未曾考虑的因素干扰的结果。

表 4 4 个立地类型样品点主要因子和立地指数观察值

样品号	持水量 (%)	N (%)	K (%)	pH	部 位	坡 度	立地指数
I 类							
14	31.5	0.29	3.6	4.9	5 (山 谷)	4 (11°-25°)	20
19	31.2	0.17	6.7	4.9	5	3 (>25°)	19
平均数 $\bar{x}$	31.4	0.23	5.2	4.9	5	3.5	19.5
I 类							
1	32.5	0.24	4.2	5.7	2 (山坡上部)	3	16
2	29.0	0.23	5.8	5.3	3 (山坡下部)	4	15
3	27.0	0.18	4.6	5.2	4 (山 脚)	3	17
10	36.6	0.23	3.4	5.0	3	3	17
16	25.1	0.17	3.3	4.9	3	3	16
20	28.8	0.19	6.9	5.0	3	3	14
35	22.5	0.15	4.8	4.9	4	3	13
平均数 $\bar{x}$	28.8	0.20	4.7	5.1	3.1	3.1	15.4
V 类							
9	32.8	0.24	1.7	5.1	1	4	7
12	29.2	0.17	2.0	4.9	2	4	8
22	31.0	0.11	1.4	4.9	3	4	12
30	25.0	0.11	1.8	4.5	2	4	9
31	34.3	0.19	1.7	4.3	3	4	7
33	39.3	0.14	1.6	5.1	3	3	10
平均数 $\bar{x}$	31.9	0.16	1.7	4.8	2.3	3.8	8.8
VI 类							
7	27.9	0.26	2.5	5.2	1 (山顶山脊)	3	6
24	32.6	0.03	0.88	4.9	2 (山坡上部)	3	13
27	29.0	0.05	1.5	5.1	1	3	7
36	23.0	0.11	7.4	4.7	1	3	9
37	27.0	0.13	2.6	4.9	1	3	10
39	27.9	0.09	6.3	4.8	1	4	8
平均数 $\bar{x}$	27.9	0.11	3.5	4.9	1.2	3.2	8.8

从表 4 中可以看出: I II 类立地属湿润肥沃酸性和微酸性土壤, 地形也较为有利 (多在山谷、山脚和阴坡, 杉木生长最好。V 类土壤含钾量低 (仅为 1.7% 左右, 含氮量一般也偏低, 杉木生长不良。第 VI 类由于地形不利, 处于山顶、山脊等开朗部位, 当风、强光照、水肥易流失, 对杉木生长也很不利。

部位、坡度、坡向是按级别计算的, 详见表 5。

表5 部 位、坡 度、和 坡 向 级 别 表

级 别	0.00	0.29	0.50	1.00	1.70	2.00	3.00	4.00	5.00
部 位	/	/	/	山顶山脊	/	山坡上部	山坡下部	山 脚	山 谷
坡 度	/	/	凸坡	/	/	/	陡坡 $>25^\circ$	中 坡 11-25°	缓 坡 $<10^\circ$
坡 向	SW (南西)	W或S (西或南)	/	SE或EW	E或N	NE (北东)	/	/	/

*坡向的编码是采用浙江亚热带林业科学研究所提出的方案。部位与坡度编码是根据样本资料绘制因子与立地指数的相关图确定的。

四、结 构 与 讨 论

(一) 本研究以与目的指标显著相关的第2、第3主组元为坐标,把广东杉木立地分成6类:〈I〉谷地多湿肥沃酸性型,平均立地指数(后略) $y_1=19.5$;〈II〉坡地肥沃酸性型, $y_2=15.4$;〈III〉坡地潮湿酸性型, $y_3=14.5$;〈IV〉陡坡中肥酸性型, $y_4=12.7$;〈V〉坡地贫钾酸性型, $y_5=8.8$;〈VI〉高亢干旱型, $y_6=8.8$ 。

(二) 聚类分析结果表明:地形条件,氮、钾水平所组成的综合指标与杉木立地指数的高低相一致。谷地阴坡、光照不过强,全氮量在0.2%以上,全钾量在4.7%以上,pH约5.0的立地杉木立地指数可达15以上;山顶山脊,山上部阳坡,全氮量0.16%以下,全钾量3.5%以下,立地指数小于10。上述结果对评价我省发展杉木生产的技术经济效益有参考价值。例如:可在严重缺钾的V类立地,增加土壤的含钾量,以提高生产力。在高亢干旱的VI类地区改种抗干旱树种。

(三) 根据本研究得出主组元方程,对未知立地作相应项目的调查后,将观测数据代入主组元方程,即可估计出该立地类属和指数,从而可以确定该地可否种杉。如根据调查某地各环境因子的数据如下表所示:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
(持水量)	(容重)	(全氮)	(全钾)	(pH)	(粉粒)	(部位)	(坡度)
36%	1克/厘米 ³	0.2%	4%	5.5	30%	4(山脚)	4(11°-25°)

将上述数据分别代入第2、第3主组元方程得:第2主组元值 $y_2=1.56$,第3主组元值 $y_3=0.51$;然后从主组元聚类图中求得它的位置在I类立地与II类立地之间(图1中上方加星号的小圆点处)。故知该地宜种杉,其立地指数估计在15以上。

(四) 主组元聚类分析利用资料的信息量大,往往以一个或两个主组元就能集中反映出原始资料的大部分信息,为多因子分析的一种重要方法。使用此法时,确定判别类的指标是一个重要环节。一般学者仅凭专业知识来判断,往往因选取指标过多,使研究工作复杂化而难于进行,或因选取的指标未进行检验而引起结果的不稳定。在选取主组

元时，一般的做法则是选取贡献联最大的一个或若干个作为分类的综合指标，但由于分类多数是有目的性的，贡献率最大的主组元不一定能提供研究所需要的分类信息，这就是分类结果往往不能达到分类目的的原因，为了克服这一缺点。刘多森^[3]（1979）等提出结合专业知识选取符合分类目的的主组元为分类的综合指标，但未曾考虑被选入的主组元是否适合分类目的的检验问题。本研究在选取分类指标时采用进步回归方法筛选与目的指标显著相关的因子；在选取主组元时，则进行主组元与目的指标的相关性检验，使分类结果达到预期目的。本研究仅用两个与目的指标显著相关的主组元作判类的依据，就能清楚地辨认出较优环境因子的水平组合（Ⅰ类与Ⅱ类立地），不仅为本省选取宜杉环境提供了科学依据，方法本身也可作为“一般数学分类”参考。

（五）聚类图中处于两类边缘的点，可用距离判别法加以调整。详见参考文献^{[1][2][4]}。由于本研究主要目的是判别宜杉立地与不宜杉立地，故未对个别边界点作调整。

参 考 文 献

- 〔1〕方开泰等：距离判别，《应用数学学报》，5（2）1982：145—154。
- 〔2〕方开泰等：聚类分析中的分解法及其应用，《应用数学学报》，5（4）1982：339—345。
- 〔3〕刘多森：主组元分析在分辨土壤类型及风化——成土过程上的应用，《土壤学报》，6（2）1979：172—182。
- 〔4〕南京大学数学系计算专业编：《概率统计基础和概率统计方法》，272—319，科学出版社，1979年。
- 〔5〕Cuanalo, de la C. H. E., Webstee, R., 1970: A comparative study of numerical classification and ordination of soil profiles in a locality near oxford. J. soil sci., 21（2）：340—352.
- 〔6〕Rao, C.R, 1973: Linear Statistical Inference and Its Applications (second edition). 590—593.

THE APPLICATION OF THE PRINCIPAL COMPONENT—CLUSTER METHOD TO IDENTIFY THE SITE TYPES OF GUANGDONG CHINA—FIRS

He Zhaoheng, Lui Youmei

Li Yingcai

(Department of Forestry)

(Department of Agricultural Machinery)

ABSTRACT

This paper deals with a simple method of the principal component-cluster to be used for classifying the site types of Guangdong China-Firs. An improvement has been made in the method of choosing classifying indexes of environmental factor, which influence the growth of the Firs eight are chosen as classifying indexes. There play an important role in the growth of the Firs. The first four bigger eigenvalue principal components in these factors, correlative matrixes are calculated. The second and third principal components, which have much to do with the site types and are of bigger contributive rate, are chosen as the integrative classifying indexes. A planar clustering graph is drawn, and then 40 samples are divided into 6 site types from which the productivity of each site type is to be judged and assessed. The results are as follows,

Types 1 and 2, which are of bigger site indexes, are good for planting China-Firs.

Types 3 and 4, which are of moderate productivity, may be planted with China-Firs.

Type 5, badly in need of potassium, should increase potassium content in the soil before increasing its productivity.

Type 6 is not good for planting China Firs because of its undesirable location.

This study provides scientific basis for estimating the economic effect of Guangdong China-Firs production and also suggests method for the systematist.