华南农业大学学报　18(3)　1997: 1~6
J. South China Agr. Univ.

华南地区籼稻米质综合评分和分类方法的研究

区靖祥[1]　伍时照[1]　甄　海[2]　吴东辉[2]　吴景强[3]

(1 华南农业大学农学系, 广州, 510642; 2 广东省农科院水稻所; 3 湛江市农科所)

摘要　衡量稻米品质的诸项指标之间并非是相互独立、同步变化的, 而是相互联系和相互制约的。这给水稻米质的评估和分类带来困难。文章运用多元因子分析的原理提出了一个对稻米品质进行综合评分和分类的方法。

关键词　稻米品质; 因子分析; 聚类分析
中图分类号　S 331

水稻是广东的主要粮食作物。近年来由于人民生活水平的提高, 发展优质稻米已经受到越来越普遍的重视。日本在 70 年代已经制定稻米品质的鉴定方法, 此法被许多研究者所采用, 并沿用至今(闵绍楷, 1981)。我国农业部也制定了鉴定稻米品质的部颁标准(中华人民共和国农业部, 1988, NY 122－86 米质测定方法)。但是由于与稻米品质有关的性状数目众多, 如何将众多的品质性状得分综合成为一个总分, 并利用它来对某地区的主要稻米品种进行评比或分类, 对当地的水稻育种和稻米市场对品种的利用和调配, 均有很大意义。多元统计中的因子分析正交旋转法为此提供了进行综合评分和分类的简单而有效的方法。本文以广东省农科院水稻所测定的一些籼稻品种的 8 个品质性状数据为例, 提出一种综合评分和分类的方法, 供同行研究者讨论参考。

1　原理和方法

目前惯用的评分方法是将测得的品质性状按一定的标准进行分级评分, 然后将所有性状的得分相加得总分, 再以此总分来对不同的品种进行比较或与确定的标准进行评定。但由于各性状之间不是独立的, 它们之间存在着不同程度的相关关系, 因此直接相加有时便显得勉强甚至误导。因子分析中的正交旋转法提供了将多元变量转化为数目较少的相互独立的综合变量(因子得分), 将这些因子得分相加, 能更合理地表现品种(关于品质)的变异。利用这些得分进行分级或分类应更为合理。

不失一般性地, 设已收集了 n 个品种的 p 个已足以充分反映稻米品质变异的变量, 记为 $X=(x_1, x_2, \cdots, x_p)$。

由于各地区的市场和消费者对"优质米"的要求有所不同, 而人们希望用一种好方法算得总分数越高的品种具有越优的质量, 因此, 分析的第一步应该是将这些变量按当地市场的喜好进行适当的转换, 使它们都变成为数值越大越好。记这 p 个转换变量为 x'_1, x'_2, …, x'_p, 然后对这 p 个新变量进行因子分析。

1997－03－12 收稿　区靖祥, 男, 53 岁, 副教授

因子分析的方法有多种，这里先采用主因子法求出主因子解，再采用方差最大旋转法寻找最适用于稻米品质评估的正交因子解。其计算过程是(张尧庭等，1983)：将数据 x'_i 标准化为平均数为0，标准差为1的标准化变量 y_i；求出向量 $Y=(y_1, y_2, \cdots, y_p)$ 的协方差矩阵(也等于相关系数矩阵)的特征值，并按从大到小顺序排列为(λ_1, λ_2, …, λ_p)后，所对应的特征向量 $U=(u_1, u_2, \cdots, u_p)$ 具有正交变换的性质。即由 U 和 Y 构成的线性组合 $Z=UY=(z_1, z_2, \cdots, z_p)$ 中，各分量 z_i 相互独立；其中第一个分量的方差为 λ_1，解释了总方差(p)中的最大部分；第二分量的方差为 λ_2，解释了总方差中的次大部分，其他分量依次递减。将这些 z_i 除以各自的标准差($\sqrt{\lambda_i}$)使之标准化，所得的 f_i 称为第 i 综合因子，习惯上保留方差贡献率之和大于85%的前 m 个综合因子，称为主要综合因子($m<p$)，简称主因子。矩阵 $A=(u_1\sqrt{\lambda_1}, u_2\sqrt{\lambda_2}, \cdots, u_p\sqrt{\lambda_p})$ 称为因子载荷阵，它的元素 a_{ij} 衡量了原变量 y_i 与主因子 f_j 之间的相关关系，载荷阵在因子分析中具有重要意义。为了寻找更适合于对稻米品质进行评估的综合变量，我们在保持变量间相互独立的前提下，对因子轴进行正交旋转，即寻找一个转换矩阵 T 并由它计算出一个新的因子载荷阵 $B=AT$，使得由 B 产生的新因子得分(新的 f_i)能更好地表现稻米的品质差异。

这些主要综合因子各衡量了影响稻米品质变异的某方面因素。当重点考察某个方面的因素时，可以用个别因素来对所有候选的品种进行排序评比，寻找在该因素上具有特别价值的品种。又由于这些综合因子都是相互独立的，所以可以将它们的值相加得综合总分。利用这些总分将全部品种排序，进行评比，便可找到综合各方面看来总分最高的品种，进一步分析便可以知道它们之所以总分最高的原因。此外还可以利用它们的个别因素得分或总分将所有品种进行分类，为亲本选育或市场开发提供理论依据。

2 实例与分析

本文以广东省农科院水稻所测定的172个籼稻品种的糙米率(x_1)、整精米率(x_2)、粒长(x_3)、粒宽(x_4)、粒长宽比($x_{3/4}$)、角质率(x_5)、直链淀粉含量(x_6)、胶稠度(x_7)和饭味品尝(x_8)等9个品质性状为例说明。由于其中的 $x_{3/4}=x_3\div x_4$，为了避免信息的重复，这里不拟采用 x_4，只采用 x_3 和 $x_{3/4}$。

首先按当地市场的喜好对拟采用的变量转换成越大越好的新变量。本例中，按广东出口(港澳)优质稻品种性状等级标准(黄超武等，1995)进行。这里认为直链淀粉含量以20%为最好，过大或过小则较差，因此令 $x'_6=1-|x_6-0.2|$；其余变量都是越大越好，无需转换，因此令 $x'_1=x_1$；$x'_2=x_2$；$x'_3=x_3$；$x'_4=x_{3/4}$；$x'_5=x_5$；令 $x'_7=x_7$；$x'_8=x_8$。

原数据和转换数据的格式如表1所示。由于篇幅所限，本文中只列出部分品种的数据。将数据标准化为 $y_1\sim y_8$，并求出它们的相关矩阵如表2所示。

又计算出相关矩阵的特征向量、特征值、方差贡献率及累计贡献率如表3所示。从表3可以看到前5个特征值之和已经超过85%，因此本例中拟保留前5个主因子。

表 **1**　**172** 个籼稻品种品质性状观测记录和转换数据(篇幅关系, 只列出部分品种)

观测顺序	品种名称	糙米率 x_1	整精米率 x_2	粒长 x_3	粒宽 x_4	粒长宽比 $x_{3/4}$	角质率 x_5	直链淀粉含量 观察记录 x_6	直链淀粉含量 转换数据	胶稠度 x_7	饭味品尝 x_8
1	七桂早	80.7	66.2	5.46	2.00	2.73	85.4	26.5	0.935	29	8
2	三二矮	80.8	57.2	5.56	2.06	2.70	79.7	26.5	0.935	28	6
3	双竹选	79.7	59.8	5.84	2.08	2.81	76.1	26.6	0.934	31	6
4	万利香	78.9	59.0	6.83	2.15	3.18	96.2	11.9	0.919	92	10
5	欧洲丝苗	79.0	54.0	5.71	2.00	2.86	82.8	27.4	0.926	30	6
…	…	…	…	…	…	…	…	…	…	…	…
170	黄壳香	75.9	68.9	5.66	1.96	2.89	88.5	18.2	0.982	45	13
171	团香谷	78.3	66.8	5.28	2.07	2.55	96.4	26.0	0.940	28	12
172	晚罗占	80.7	67.5	6.73	1.82	3.70	97.2	18.3	0.983	58	15

表 **2**　**8** 个性状的相关系数矩阵

变量	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
y_1	1.000 0	0.270 7	0.309 7	0.131 6	0.001 4	−0.054 2	0.054 1	0.068 7
y_2	0.270 7	1.000 0	0.009 7	0.226 9	0.459 9	0.102 1	0.038 5	0.151 9
y_3	0.309 7	0.009 7	1.000 0	0.651 3	0.145 9	0.136 4	0.310 3	0.275 9
y_4	0.131 6	0.226 9	0.651 3	1.000 0	0.436 0	0.141 7	0.407 2	0.445 2
y_5	0.001 4	0.459 9	0.145 9	0.436 0	1.000 0	0.190 6	0.277 3	0.429 5
y_6	−0.054 2	0.102 1	0.136 4	0.141 7	0.190 6	1.000 0	0.383 4	0.226 2
y_7	0.054 1	0.038 5	0.310 3	0.407 2	0.277 3	0.383 4	1.000 0	0.444 8
y_8	0.068 7	0.151 9	0.275 9	0.445 2	0.429 5	0.226 2	0.444 8	1.000 0

表 **3**　相关矩阵的特征向量、特征值、方差贡献率及累计贡献率

变量	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8
y_1	0.154 8	−0.629 9	−0.247 5	0.500 2	0.345 4	0.056 6	0.298 5	0.237 1
y_2	0.239 9	−0.469 7	0.555 9	0.167 4	−0.168 7	−0.142 7	−0.534 2	−0.224 8
y_3	0.378 7	−0.143 9	−0.556 5	−0.071 5	−0.352 7	0.128 6	0.031 3	−0.615 8
y_4	0.478 9	−0.078 8	−0.205 3	−0.332 2	−0.305 1	−0.071 8	−0.206 7	0.686 3
y_5	0.389 9	−0.054 7	0.490 0	−0.278 0	−0.073 6	−0.038 2	0.716 3	−0.087 3
y_6	0.249 9	0.439 6	0.143 3	0.669 8	−0.371 2	0.342 3	0.055 4	0.130 7
y_7	0.394 7	0.361 2	−0.109 5	0.217 5	0.323 4	−0.732 1	−0.047 0	−0.108 4
y_8	0.418 8	0.170 3	0.072 5	−0.182 0	0.622 1	0.547 9	−0.251 9	−0.081 2
特征根值	2.825 0	1.277 4	1.229 9	0.884 1	0.667 6	0.491 0	0.383 7	0.241 2
方差贡献	0.353 1	0.159 7	0.153 7	0.110 5	0.083 5	0.061 4	0.048 0	0.030 2
累计贡献	0.353 1	0.512 8	0.666 5	0.777 1	0.860 5	0.921 9	0.969 8	1.000 0

表4 经10次方差最大旋转后的因子载荷阵

变量	f_1	f_2	f_3	f_4	f_5
y_1	0.144 7	0.095 4	0.035 9	−0.049 3	0.957 6
y_2	−0.011 5	0.882 0	−0.052 0	0.082 8	0.284 2
y_3	0.908 1	−0.072 4	0.105 4	0.094 9	0.223 7
y_4	0.832 8	0.285 3	0.300 4	0.022 7	−0.041 1
y_5	0.201 6	0.763 4	0.370 1	0.040 3	−0.196 3
y_6	0.054 4	0.100 7	0.128 8	0.951 0	−0.055 1
y_7	0.231 5	−0.047 3	0.694 4	0.438 2	0.056 1
y_8	0.167 4	0.201 4	0.870 6	0.001 4	0.012 9

表4列出了经10次方差最大旋转后的因子载荷阵。从表4中可以看到，第1主因子 f_1 与粒长 y_3 和粒长宽比 y_4 有较大的正相关，而与其他原变量的相关较小，因此可以认为 f_1 是粒型品质因子，即 f_1 大的品种粒型较好，反之较差；第2主因子 f_2 与整精米率 y_2 和角质率 y_5 有较大正相关，而与其他原变量的相关较小，因此可以认为 f_2 是碾磨品质因子，即 f_2 大的品种不易脆折，碾磨性较好；第3主因子 f_3 与饭味品尝 y_8 和胶稠度 y_7 有较大正相关，而与其他原变量的相关较小，因此可以认为 f_3 是蒸煮品质因子，即 f_3 大的品种的蒸煮品质较好，反之较差；第4主因子 f_4 与直链淀粉含量 y_6 和胶稠度 y_7 有较大正相关，而与其他原变量的相关较小，因此可以认为 f_4 是软硬品质因子，即 f_4 大的品种的软硬口感较好，反之较差；第5主因子 f_5 与糙米率 y_1 有较大正相关，而与其他原变量的相关较小，因此可以认为 f_5 是出米量品质因子，即 f_5 大的品种的出米量较高，反之较低。

表5列出了部分品种的主因子得分和名次，如果要选择粒型好的品种，可观察 f_1 的名次。例如172号品种（晚罗占）的粒型就比列出的其它品种好，其次的是4号品种万利香。如果要选择碾磨性质好的品种，可观察 f_2 的名次。例如171号品种（团香谷）的碾磨综合性质比列出的其它品种好，其次的是170号和172号。f_3、f_4 和 f_5 也可类似地作为对单因子进行选择的依据。表5的因子总分是前5个主因子得分的和，可以作为综合评估的依据。例如在所列的8个品种中，172号品种的因子总分最高，可以认为从综合情况看，它比表列的其它7个品种都好。从表5的最下一行可以看到172号品种之所以总分最高的原因是它的第3主因子（蒸煮品质）和第1主因子（粒型品质）都名列前矛（分别排在17和23名），而其它因子的名次也不落后。而其它品种，例如4号品种，虽然第3主因子名次较前，但其它方面却比较落后，它的第2主因子（碾磨品质）和第4主因子（软硬品质）都排在140名以后。

根据主因子得分可以利用系统聚类或动态聚类方法对品种进行分类。前法要求较多的计算机内存，如果计算机内存不够，可采用后法。图1是根据前5个主因子的得分用欧几里德距离和最小距离法对本文列出的8个品种进行聚类所得的分类树型图，横坐标是品种或类别间合并的欧氏距离，纵坐标是品种编号，将表5的数据和图1的显示比较，可以看到最先结合为一类的是2号品种和3号品种，它们的第2和第3主因子得分和名次都很相近，第3和第4主因子得分和名次也相差不远；第二步是将1号品种加入上面得到的类别中；第三步又加入5号品种，之后加入171号品种；继而加入170号品种。第4号品种与第172号品种两者间比较相近但与上述其它品种都距离较远；如果在欧氏距离为2.2处切开，可以将这8个品种分为三类（即第1、2、3、5、170、171号品种为一类，第4号和第172号品种各为一类）；如果在欧氏距离为

2.4 处切开, 可以将这 8 个品种分为两类(即第 1、2、3、5、170、171 号品种为一类, 第 4 号和第 172 号品种合为另一类)。从表 5 可以看到后两品种的品质得分和名次都比前面品种的高。图 2 是根据前 5 个主因子的得分用欧几里德距离和最小距离法对全部 172 个品种进行聚类所得的分类树型图, 由于品种数较多, 这里将图形旋转了 90 度, 即横坐标表示品种的编号(由于篇幅所限, 品种号未能标出), 纵坐标标出品种或类别间合并的欧氏距离, 该图清楚地显示出整个聚类的过程和结果。由于本文没有列出全部 172 个品种的数据, 这里也不拟对分类过程和分类结果作详细讨论。仅用图形为读者提供一个总的印象。

表 **5**　各品种经 **10** 次旋转后的主因子得分和名次(篇幅关系, 只列出部分品种)

品种	第 1 主因子		第 2 主因子		第 3 主因子		第 4 主因子		第 5 主因子		因子总分	
编号	得分	名次	得分	名次	得分	名次	得分	名次	得分	名次	得分	名次
1	−0.975 7	145	0.288 9	82	−0.110 5	83	−0.668 5	115	0.764 4	33	−0.701 4	105
2	−0.507 3	115	−0.701 5	140	−1.033 9	152	−0.365 6	98	0.639 0	44	−1.969 3	144
3	−0.028 3	83	−0.703 5	141	−1.236 1	155	−0.297 4	91	0.403 4	66	−1.861 8	141
4	0.937 3	32	−0.776 6	146	2.210 1	5	−1.034 1	144	0.092 8	88	1.429 4	43
5	−0.281 1	97	−0.891 4	151	0.127 1	72	−1.121 2	149	−0.098 6	107	−2.265 1	153
…	…	…	…	…	…	…	…	…	…	…	…	
170	−0.867 2	142	0.535 8	55	0.935 0	33	0.905 0	40	−0.788 8	138	0.719 8	59
171	−1.625 5	165	0.821 7	27	1.016 3	27	−0.889 2	130	−0.165 8	113	−0.842 6	109
172	1.204 7	23	0.508 9	59	1.451 0	17	0.691 4	53	0.490 8	57	4.346 9	5

本文的统计分析在国际通用统计软件 Statistica 下运行。

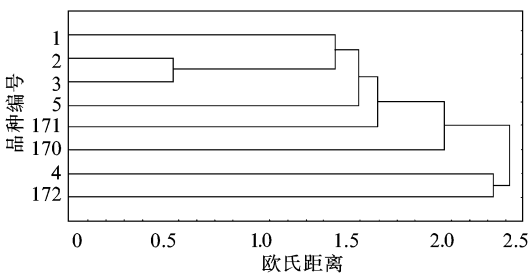

图 1　根据 5 个主因子得分用欧氏距离和最小距离法对本文列出的 8 个品种进行分类的树型图

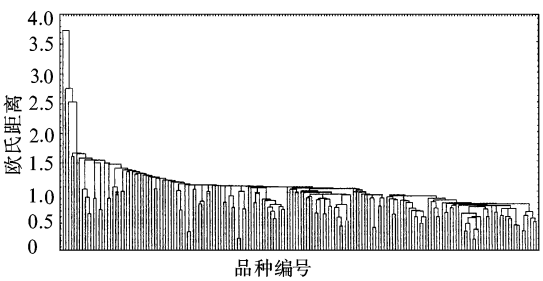

图 2　根据 5 个主因子得分用欧氏距离和最小距离法对全部测定的 172 个品种进行分类的树型图

3　讨论和结论

因子分析已广泛应用来对样本进行排序和分类。本文将其改进应用于对水稻品质性状进行评比和分类。

由于各地区、民族、年龄的人对“优质米”的定义不同, 因此评估的方法应能具有一定的弹性。本法采用将全部原变量 x_i 转换为越大越好的新变量 x'_i 的方法, 简单易行, 例如, 如

果某地区的人比较喜欢"糯性"的大米,可以将最佳的直链淀粉含量从本例的20%降低到适当的水平。如果资料中有一个性状 x_i(如垩白粒率)是越小越好的,则可以将其转换为 $x'_i = 100(\%) - x_i$ 或 $x'_i = 1/x_i$。

衡量稻米品质的性状当然不只是本例中8个。为了使评估结果更能反应实际情况,应收集一切与稻米品质有关的性状(如垩白粒率、蛋白质含量、某种氨基酸的含量等),参与分析。收集的性状越齐全,分析结果将越可靠。

饭味品尝评分则较多地受人为因素的影响,带主观性且误差较大。目前不少学者正开展对此性状与其他客观性状间的相关关系的研究(刘宜柏等,1989;黄超武等,1990),如果这些研究成功,可望用其它性状的数据来代替或间接测定它,分析结果可能会更加客观。

本例鉴定了172个籼稻品种。在实际应用中,应尽可能收集齐全该地区的全部品种,建立完整的数据库,进行分析。品种越齐全,分析的结果将越有说服力。

只要品种数目齐全,性状项目充分,此法也可应用于其它地区的其它水稻类型的稻米品质评比和分类工作中。

参 考 文 献

刘宜柏,黄金英. 1989. 稻米食味品质的相关性研究. 江西农业大学学报, 11(4):1~5

闵绍楷编译. 1981. 稻米品质的鉴定和改良. 国外农学—水稻, (2):113~123

张尧庭,方开泰. 1983. 多元统计分析引论. 北京:科学出版社, 328~338

黄超武,黄远生. 1990. 谷物作物品质性状遗传研究进展. 南京:江苏科技出版社, 58~63

黄超武,伍时照. 1995. 广东出口(港澳)优质稻品种性状等级标准. 见:黄超武主编. 水稻品种种性研究. 广州:广东科技出版社, 109~110

A STUDY ON COMPREHENSIVE EVALUATION AND CLASSIFICATION METHOD FOR INDICA RICE QUALITY IN SOUTH CHINA REGIONS

Ou Jingxiang[1] Wu Shizhao[1] Zhen Hai[2] Wu Donghui[2] Wu Jingqiang[3]

(1 Dept. of Agronomy, South China Agr. Univ., Guangzhou, 510642;
2 Rice Research Institute, Guangdong Academy of Agr. Sci.;
3 Zhanjiang Agr. Research Institute)

Abstract

The indices used to evaluate rice quality are interdependent as well as changeable synchronously. There exist certain correlation and association among them. This makes the evaluation and classification of rice quality difficult. The results, based on factor analysis in multi—variate statistics, suggested a method which can be used to evaluate or classify rice varieties concerning their quality.

Key words rice quality; factor analysis; clustering