在本色谱分离条件下，维生素α、γ、δ-E均能得到较好的分离，整个分离过程仅需10min。（见图1）

2.2 方法回收率实验

取一定油样加入标准品，用HPLC测定、回收率见表2。

表2 回收率测定结果

名称	α-V_E	γ-V_E	δ-V_E
实测量（ng）	20	18	15
加标量（ng）	15	15	15
总量（ng）	33	31	27.8
回收率（%）	64.2	93.9	92.7

2.3 最低检出浓度与精密度实验

最小检出限以2倍噪音计，本法最小检出浓度为0.4μg/ml，取5μl维生素E标准液连续进样6次，结果见表3。

表3 精密度实验结果

样品名称	测定次数	实测范围	变异系数（%）
α-V_E	6	19.01～20.56	2.55
γ-V_E	6	19.25～20.62	2.17
δ-V_E	6	19.11～20.75	2.75

2.4 维生素E又称生育酚，英文名Tocopherol，属酚类化合物，由于其结构苯环上连接的甲基的位置和个数不同，而分为α、β、γ、δ等同系物，其中以α-维生素E的生物活性最强。维生素E在生物体内主要表现为抗氧化作用，因此被称为消除自由基的营养素，对预防衰老与健康有着密切的关系。基于不同的植物油所含的维生素E的种类和含量不尽相同，我们采用反相高效液相法对福州地区市售的主要食用植物油中各种异构体进行分离测定研究（见表4）

表4 植物油中维生素E各异构体的含量

样品	维生素$E_{总}$ (mg/kg)	α-V_E (mg/kg)	γ-V_E (mg/kg)	δ-V_E (mg/kg)
核桃油	357.5	6.6	312	38.9
玫瑰茄籽油	347.5	22.4	310.6	14.5
香麻油	324.4	1.6	313.6	9.2
花生油	102.7	8.2	87.1	7.4
葵籽油	136.5	95.9	33.5	7.1
棕榈油	38.1	31.6	6.6	0.1

从表4中所列各种食用植物油分析结果看出，不同品种食用植物油中维生素E的含量差异极显著，测定结果的均值顺序为：核桃油＞玫瑰茄籽油＞香麻油＞葵籽油＞花生油＞棕榈油。值得我们做进一步研究的是：实验选自的核桃油不需再经任何烹调就可直接食用的软胶囊包装产品，避免了其中维生素E在高温过程中的损耗，这给我们开发新优产品提供了有益的提示，同时这种含维生素E极其丰富的产品与人体健康长寿之间的关系都有待于做深入的研究（核桃油原料来自格鲁吉亚共和国，那里的人们经常食用核桃，平均寿命较长）。此外不同植物油维生素E各异构体的含量具有特征性，对它们各自的差异很容易用高效液相色谱法进行辨别，这为鉴定鉴别和监督管理提供了可靠的技术保证。作为维生素E摄取的主要来源之一，植物油中维生素E含量以上测定结果可作为评价本地区群体维生素E摄入量的参考，也为合理搭配膳食提供依据。

参考文献

1 中国预防医学科学院标准处编.食品卫生国家标准汇编(2).北京：中国标准出版社，1992，9.

2 何照范等编著.保健食品化学及其检测技术.中国轻工业出版社.第一版.北京：1998，5.

中国白酒香型的化学模式识别（II）
——聚类分析

陈华 郁志勇 朱国斌 中国人民大学商品学系 北京 100872

摘 要 以46种白酒为样本，以每个样本的乙酸乙酯、己酸乙酯等17种香味成分的气相色谱、气质联用定性定量分析数据和感官评价结果为指标，用聚类分析研究了白酒香型数据。从统计学的观点来看，本文所选聚类方法可行，所得分析结果理想。

关键词 聚类分析 香型 白酒

Abstract Based on the data of 17 flavor components in 46 samples that were obtained by GC,GC-MS and sensory evaluation the cluster analysis was used in the aroma recognition of the samples.The correlation between sensory analysis and chemical compositions was disclosed.Viewing from a statistical perspective,the results were satisfactory.

Key words Cluster analysis Aroma White liquor

1 原理[1～6]

聚类分析是研究分类问题的一种多元统计分析方法，在很多学科领域都有着广泛的应用。由于它不是根据事先的定义而是根据数据本身来对样本（变量）分组，以达到降维的目的，这是一种“无师可学”的分类方法，其分类原则可概述为“物以类聚”。最常用的聚类方法为系统聚类法，其基本思想是认为所研究的样本或指标间存在程度不同的相似性（亲疏关系），于是根据一批样本的多个观测指标，具体地找出一些能度量样本或指标间相似程度的统计量，并以这些统计量为划分类型的依据，把一些性质相似程度较大的样本聚为一类，而把另一些彼此间相似程度较大的样本聚为另一类，把关系密切的聚合到一个小的分类单位，把关系疏远的聚合到一个大的分类单位，直至所有样本聚合完毕，所有类型一一划分出来，形成一个由小到大的分类系统，并得到一张能表示所有样本（或指标）间亲疏关系的谱系图。

常用的聚类统计量有距离系数和相似系数两种，从本文的研究目的出发，最终选用距离系数为聚类统计量，并以系统聚类法为最终的分析方法。

2 结果与讨论

2.1 用原始数据进行聚类分析的结果与讨论

为了得到最好的聚类效果，本研究中分别尝试了不同的方法。比较这11种分析方法所得的结果可知：错判率最小的是最大距离法所得的结果（4个错判，清36、44、45、46号被误判为酱香型样本，错判率8.69%），其次为类平均法、可变类平均法、McQuitty’s相似分析法、中间距离法、重心法、最大似然估计法、Ward最小方差法所得的结果（各自分别有6个样本被误判：浓34、35号样本，清36、44、45、46号均被误判为酱香型样本），而最短距离法、密度估计法、两阶段密度估计法所得的结果不理想。故最终选用的聚类方法为最大距离法，用此方法进行聚类分析，各步的聚类情况详见表1，最后所得的系统聚类图详见图1：

为了更好地了解感官分析结果与各香型酒间化学成分的相关性，也为了更好地理解各类数与各香型间的关系，下面将对表1中的各结果做详细分析：

2.1.1 由半偏R^2的值可见，其在刚开始时一直很小，直至最后两步才发生数量级的突变，这意味着在聚类之初，合并两个类时产生的方差减少比例缓慢地变化，每一步合并对信息的损失程度都较小，直至最后两步这种信息损失才发生实质变化，因此，半偏R^2支持将原始数据分阶段3类。

同理，由于反映累计聚类结果的平方多重相关的R^2值由3分类的0.692014急剧下降为2分类时的0.328405，故上一次聚类的R^2值减去本次半偏R^2等于本次R^2也支持将原始数据分为3类。类似地，在均匀零假设下R^2的近似值期望值也支持3分类。

2.1.2 尽管度量在当前水平下所有类的分类程度的伪F值（Pseudo F）在3、14分类处都有峰值，但相对而言，在3分类的峰值更大，故伪F值支持3分类。

2.1.3 虽然度量最近合并的两个类间分离程度的伪t^2值（Pseudo t^2）在2分类处的峰值最大，但由于该值所支持的最佳分类数为出现最大峰值的前一行所对应的分类数，故该值也支持3分类。

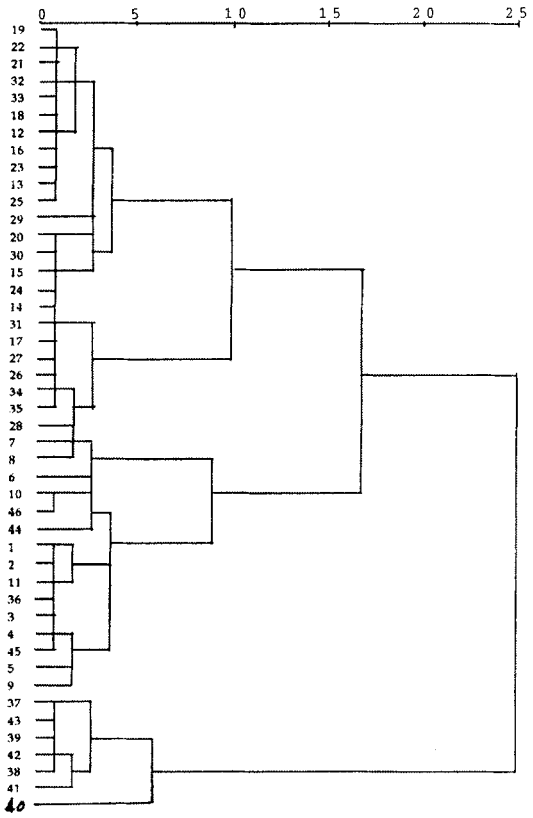

图1 系统聚类图

2.1.4 对立方聚类标准CCC（Cubic Clustering Criterion）进行类似分析可知该标准认为将原始数据分为2类较好。

2.1.5 分析图1和表1中的相关信息可见：除清36、44、45、46号样本被误分为酱香型样本外，其余样本的香型分类均完全正确。

综合上述结果可知：同一香型酒的数据间确实存在相似性，而不同香型酒的数据间相似性则不大；根据这些相似性，从各香型酒的化学组成数据出发，能实现不同香型酒的分类。同时，这些聚类分析结果也说明白酒样本的化学成分与来自感官评价的香型之间有相关性。

但由于CCC标准只支持将原始数据分为2类，这就意味着原有样本并未被彻底分开，观察图1和表1中的结果可知：是酱香型样本和清香型样本没有完全分开（4个清香型样本被误分为酱香型样本）。分析这4个误分样本的成因，我们认为主要有以下几点：

1、来自聚类原理的影响。由于统计方法的实质是给出某一置信度下，某一结果落在某一置信区间的概率为多少，这一方法本身是允许误分样本存在的。另外，在进行聚类分析时，由于各集合间存在着“桥”或散逸，分类到不同类别中去的“样本对”可能比在同一集合中的某些“样本对”有更多的相似性[4]等原因的存在，也可能导致误分类。

表1 聚类结果汇总表

类序号	被合并的类		新类中的观测数	半偏 R^2	R^2	R^2 的近似期望	CCC 立方聚类判据	伪 F 值	伪 t^2 值	正规化最大距离
45	OB19	OB22	2	0.000092	0.999908	.	.	247.60	.	0.070820
44	OB32	OB33	2	0.000169	0.999739	.	.	178.03		0.096214
43	OB17	OB27	2	0.000371	0.999368	.	.	112.94	.	0.142364
42	OB14	OB31	2	0.000459	0.998909	.	.	98.33	.	0.158353
41	OB12	OB16	2	0.000497	0.998413	.	.	78.62	.	0.164719
40	OB39	OB42	2	0.000688	0.997725	.	.	67.46	.	0.193861
39	CL43	OB26	3	0.000827	0.996898	.	.	59.19	2.23	0.199852
38	OB3	OB4	2	0.000788	0.996109	.	.	55.35	.	0.207571
37	OB18	CL44	3	0.000923	0.995186	.	.	51.69	5.45	0.207755
36	CL41	PB23	3	0.001179	0.994007	.	.	47.39	2.37	0.237085
35	OB38	CL40	3	0.001164	0.992844	.	.	44.89	1.69	0.239519
34	OB1	OB2	2	0.001315	0.991529	.	.	42.56	.	0.268057
33	OB34	OB35	2	0.001328	0.990201	.	.	41.05	.	0.269355
32	OB13	OB25	2	0.001340	0.988861	.	.	40.09	.	0.270620
31	CL42	CL39	5	0.001461	0.987401	.	.	39.18	2.64	0.3087340
30	OB11	OB36	2	0.001732	0.985669	.	.	37.95	.	0.307658
29	OB20	OB30	2	0.001745	0.983923	.	.	37.16	.	0.308813
28	OB10	OB46	2	0.001870	0.982054	.	.	36.48	.	0.319647
27	OB37	OB43	2	0.002162	0.979892	.	.	35.61	.	0343713
26	CL36	CL32	5	0.001842	0.978050	.	.	35.65	1.83	0.345474
25	CL26	CL37	8	0.003660	0.974390	.	.	33.29	3.69	0.369615
24	CL38	OB45	3	0.002638	0.971752	.	.	32.91	3.35	0.386312
23	OB15	OB24	2	0.003151	0.968601	.	.	32.25	.	0.414940
22	CL45	OB21	3	0.003995	0.964606	.	.	31.15	43.53	0.415699
21	CL34	CL30	4	0.004445	0.960161	.	.	30.13	2.92	0.475011
20	OB5	OB9	2	0.004148	0.956013	.	.	29.74	.	0.476125
19	OB7	OB8	2	0.004164	0.951849	.	.	29.65	.	0.477005
18	OB28	CL33	3	0.005798	0.946051	.	.	28.88	4.37	0.508157
17	CL24	CL20	5	0.006828	0.939223	.	.	28.01	2.70	0.529725
16	CL25	CL22	11	0.009694	0.929529	.	.	26.38	6.37	0.554946
15	CL35	OB41	4	0.006354	0.923175	.	.	26.61	6.86	0.556159
14	CL23	C;29	4	0.006668	0.916506	.	.	27.02	2.72	0.639911
13	CL28	OB44	3	0.009432	0.907075	.	.	26.84	5.04	0.689944
12	CL27	CL15	6	0.007826	0.899249	.	.	27.59	3.02	0.700611
11	CL16	OB29	12	0.007333	0.891915	.	.	28.88	3.14	0.704011
10	CL31	CL18	8	0.013303	0.878612	.	.	28.95	7.97	0.754602
9	OB6	CL18	3	0.010994	0.867618	0.857827	0.7114	30.31	2.64	0.762279
8	CL11	CL14	16	0.015189	0.852428	0.840823	0.7732	31.36	5.03	0.791735
7	CL21	CL17	9	0.015533	0.836896	0.820359	1.0158	33.35	4.97	0.849346
6	CL7	CL13	12	0.013594	0.823301	0.795007	1.6223	37.27	2.79	0.857962
5	CL12	OB40	7	0.018120	0.805181	0.762332	2.2835	42.36	4.98	1.037405
4	CL6	CL9	15	0.039286	0.765896	0.716678	1.9270	45.80	6.59	1.287739
3	CL8	CL10	24	0.073882	0.692014	0.643729	1.6726	48.31	20.06	1.364473
2	CL4	CL3	39	0.363608	0.328405	0.497749	-3.2109	21.52	49.5	1.812468
1	CL2	CL5	46	0.328405	0.000000	0.000000	0.000	.	21.52	.202684

2、来自样本量的影响。尽管单个样本的分布具有随机性，但大量样本的分布却遵循一定的统计规律，样本量越大，用统计方法所得的结果就越准确。而本研究中，所用数据由于受客观条件的限制，其样本量不太大（共46个样本，其中：浓香型样本24个，清香和酱香型样本各11个）。由于样本量较少，尤其是清香和酱香型样本量的不足，使得这些数据的分布并不能完全代表各香型数据的总体分布，并最终导致酱香和清香样本间的误判。

3、来自原始信息本身的影响。由于原始数据中所有样本香型的判定都来自感官分原，而感官分析存在很大的主观性，我们不能100%地保证这些香型数据的正确性。一旦各样本的感官评价结果有误，它们会直接导致误分类的产生。而比较我

们用不同方法和数据所得的各结果可知：清45、46号样本始终被误判为酱香型，这使我们不能不对这两个样本香型的可信度打一折扣。

4、来自所分析数据的影响。尽管决定白酒风味的物质有三种：色谱骨架成分、复杂成分和协调成分[29]，但本研究由于受资料来源的限制，只对色谱骨架成分方面的有关数据进行了分析，而信息资料的不全也会导致误判的产生。

5、来自聚类方法的影响。本文所用的聚类方法为最大距离法，但最大距离法却有使聚类结果严重倾向于产生直径粗略相等的类和聚类值可能被异常值严重扭曲的缺陷。而本研究中却不能保证所收集到的数据中不含异常值，一旦有异常值存在，这也会直接导致样本的误分类。

2.2 用因子得分进行聚类分析的结果与讨论

为了进一步简化聚类计算，考虑用降维后的因子得分数据进行聚类分析。为了保证用因子得分数据进行聚类分析的结果与用原始数据进行聚类分析的结果有可比性，也为了更好地比较两者的分析效果，用因子得分数据进行聚类分析时采用了和用原始数据进行分析时相同的程序和方法，此次聚类过程中，各步的聚类情况详见表2。最后所得的系统聚类图详见图2。

同理1，对表2中的结果进行如下分析：

2.2.1 由半偏R^2值可知：开始时，每一步的损失都较小，变化也缓慢，直至最后一步，信息损失才发生数量级的突变，故半偏R^2支持将因子得分数据分为2类。

2.2.2 由于R^2值只有在分类数由2变为1时才发生急剧变化，故R^2值也是支持2分类。同理，在均匀零假设下R^2的近似期望值也支持2个分类。

2.2.3 由于立方聚类标准CCC在2分类处有峰值，故该标准支持2分类。

2.2.4 由于伪F值在2分类处有峰值，伪t^2值在1分类处有最大峰值，故这两个统计量支持2分类。

由此可见，用因子得分数据进行聚类分析时，所有的统计量都支持将原有样本分为两类，这就意味着用因子得分数据进行分析时，原有类间并没有完全分开，观察图2和表2中的结果可知：主要还是酱香型样本和清香型样本没有完全分开。而观察此次聚类过程中的误判情况可知：酱香型样本全部被判对，浓香型样本有两个被误判（浓34、35号被误判为酱香型），清香型样本有两个被误判（清45、46号被误判为酱香型），总误判数为4，误判率8.69%。将这一结果和前面的结果相比较，我们发现：虽然在这两种结果中的误判数均为4，但却有两个样本在判别上有差异，分析这些差异的成因、以及所支持的分类数减少的原因，我们认为除了前述5点原因外，还应考虑以下2点：

1）由于进行因子分析时，只保留了5个主因子，而这5个主因子所表述的累计变异比只有94.20%，因此，用这些损失了部分原始信息的综合指标进行分析，其结果难免会有差异。同时由于所选取的样本量不够大，数据所代表的信息就会有偶然性，再加上所选的分析指标不够全面，指标间的相似性较大（都为酯类和醇类指标），所代表的信息量本身就不够全面，因此，用进一步损失后的信息来进行聚类分析，最终导致了与理想结果的偏离。

2）由于因子分析中，因子得分是通过回归方法得到的，故其值都是估计值，用不精确的结果进行分类，也难免会产生对分析结果不利的影响。

尽管表2中列出的各指标都认为将样本分为两类为定，但在实际过程中对聚类类别的最终确定还要由所研究的问题来决定[3~5]，而在本研究中，由于已知香型有三类，通过重选分类界限也可将原有样本分为3类，且分类结果较理性（误分样本数为4，误分率8.69%），故用因子得分数据进行香型的分类也是可行的。

3 结论

3.1 白酒香型的感官评价结果和感官评价结果与仪器分析数据间有相关性。同类香型酒的数据间有相似性，而不同香型酒间的数据间不相似。

3.2 用聚类的方法不论是分析各样本化学组成的原始数据还是因子得分数据，都可实现各样本香型的分类。

3.3 由于用原始数据进行聚类分析的效果和用因子得分数据进行分析的效果相差不大，这也说明以前所做的因子分析是成功的，由该分析所提出的这5个主因子已代表了17种原始信息中的绝大部分内容。另外，虽然后者相对于前者来说，分析效果并无什么改进，甚至可认为其效果略逊于前者，但由于用这些简化的因子得分数据进行聚类分析具有使工作量大幅度降低的优点，这种分析途径仍是可取的。

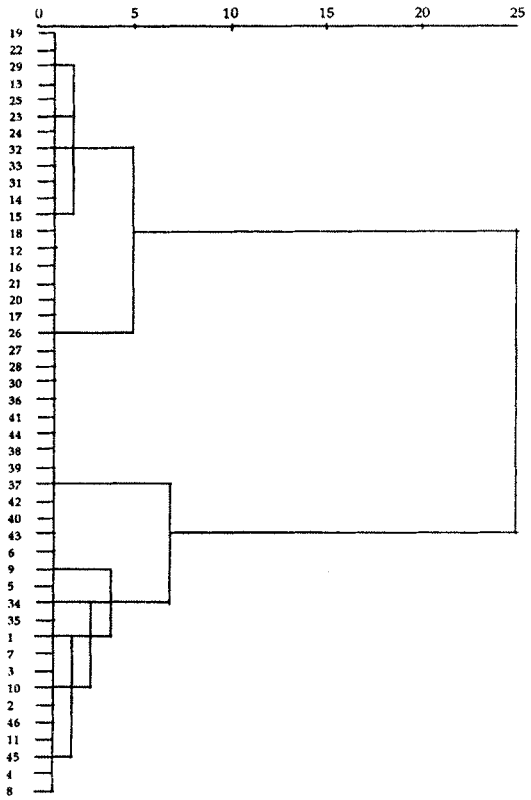

图2 系统聚类图

表2 聚类结果汇总表

类序号	被合并的类		新类中的观测数	半偏 R^2	R^2	R^2 的近似期望	CCC 立方聚类判据	伪 F 值	伪 t^2 值	正规化最大距离
45	OB19	OB22	2	0.000006	0.999994	.	.	3992.01	.	1.2883
44	OB32	OB33	2	0.000019	0.999975	.	.	1850.10	.	2.3810
43	OB3	OB10	2	0.000034	0.999941	.	.	1200.94	.	3.1638
42	OB36	OB41	2	0.000105	0.999835	.	.	592.37	.	5.5380
41	OB38	OB39	2	0.000107	0.999728	.	.	460.15	.	5.5828
40	OB13	OB25	2	0.000124	0.999605	.	.	389.07	.	6.0049
39	OB2	OB46	2	0.000137	0.999468	.	.	345.76	.	6.3252
38	OB27	OB28	2	0.000158	0.999309	.	.	312.89	.	6.7883
37	OB17	OB26	2	0.000180	0.999130	.	.	286.95	.	7.2426
36	OB31	CL44	3	0.000226	0.998904	.	.	260.33	11.61	7.3370
35	OB6	OB9	2	0.000225	0.998678	.	.	244.48	.	8.1047
34	OB12	OB16	2	0.000226	0.998452	.	.	234.61	.	8.1168
33	OB14	CL36	4	0.000283	0.998170	.	.	221.57	2.30	8.3834
32	CL45	OB29	3	0.000343	0.997827	.	.	207.41	60.17	9.0435
31	CL39	CL43	4	0.000429	0.997398	.	.	191.66	5.00	9.0673
30	OB37	OB42	2	0.000308	0.997090	.	.	189.01	.	9.4829
29	CL38	OB30	3	0.000355	0.996734	.	.	185.29	2.25	10.4631
28	OB15	OB18	2	0.000421	0.996314	.	.	180.17	.	11.0727
27	CL30	CL41	4	0.000429	0.995885	.	.	176.84	2.06	11.5610
26	CL34	OB21	3	0.000413	0.995472	.	.	175.88	1.83	12.0931
25	CL42	OB44	3	0.000424	0.995048	.	.	175.82	4.03	12.1791
24	CL31	OB11	5	0.000520	0.994528	.	.	173.84	2.60	13.1877
23	CL40	CL32	5	0.001158	0.993370	.	.	156.64	7.36	13.5744
22	OB34	OB35	2	0.000643	0.992727	.	.	155.99	.	13.6953
21	CL37	CL29	5	0.000628	0.992099	.	.	156.96	2.72	15.6421
20	CL26	OB20	4	0.001023	0.991076	.	.	151.98	3.20	16.0994
19	CL27	OB40	5	0.001112	0.989964	.	.	147.96	3.95	16.7243
18	CL24	OB45	6	0.001228	0.988736	.	.	144.58	4.38	16.9775
17	OB5	CL35	3	0.000838	0.987898	.	.	147.95	3.72	17.3389
16	CL20	CL28	6	0.000813	0.987085	.	.	152.86	1.56	19.1862
15	CL18	OB4	7	0.001399	0.985686	.	.	152.48	2.98	19.5714
14	OB1	OB7	2	0.001408	0.984278	.	.	154.11	.	202581
13	OB23	OB24	2	0.001483	0.982795	.	.	157.09	.	20.7945
12	CL25	CL19	8	0.002622	0.980173	.	.	152.80	6.33	22.8151
11	CL15	OB8	8	0.001371	0.978802	.	.	161.61	2.20	24.7500
10	CL12	OB43	9	0.002251	0.976551	.	.	166.58	3.08	24.7866
9	CL16	CL33	10	0.004336	0.972215	0.948375	4.4720	161.83	10.13	27.8490
8	CL14	CL11	10	0.003987	0.968228	0.940487	4.5721	165.43	4.89	32.6452
7	CL23	CL13	7	0.006013	0.962215	0.930447	4.5066	165.53	9.66	34.2181
6	CL9	CL7	17	0.011426	0.950789	0.971183	3.9111	154.57	10.15	41.2251
5	CL8	CL22	12	0.010700	0.940089	0.898761	4.0390	160.84	9.59	46.4901
4	CL5	CL17	15	0.023380	0.916710	0.871306	3.4767	154.09	13.26	59.6454
3	CL6	CL21	22	0.041462	0.875247	0.823498	2.0978	150.84	27.99	67.4251
2	CL4	CL10	24	0.050276	0.824971	0.692649	4.1883	207.39	20.61	75.9612
1	CL2	CL3	46	0.824971	0.000000	0.000000	0.0000	.	207.39	152.6355

参考文献

1 王学仁，王松桂.实用多元统计分析.上海：上海科学技术出版社，1990.

2 何晓群.现代统计分析方法与应用.北京：中国人民大学出版社，1999，第二版.

3 董大钧.SAS-统计分析软件应用指南.北京：电子工业出版社，1993.

4 高惠璇等编译.SAS系统-SAS/STAT软件使用手册.北京：中国统计出版社，1992.

5 卢纹岱，金水高.SAS/PC统计分析软件实用技术.北京：国防工业出版社，1996.

6 董大钧，张尔强，何武等译.[美]SAS研究所著.SAS统计过程指导.沈阳：辽宁科学技术出版社，1992.