第26卷 第4期
2007年11月

人 类 学 学 报
ACTA ANTHROPOLOGICA SINICA

Vol.26, No.4
Nov., 2007

聚类分析和主成分分析方法在人类学研究中价值的判定

吴秀杰[1]，张全超[1,2]，李海军[1,3]

（1. 中国科学院古脊椎动物与古人类研究所，北京 100044；2. 吉林大学边疆考古研究中心，长春 130012；
3. 中国科学院研究生院，北京 100049）

摘要：本文以生活在不同地区的9组人群的成年男性头骨（668例）为主要研究对象，通过对其14项测量性状的聚类分析和主成分分析，探讨多变量统计分析方法在人类学研究中的价值。结果显示：欧氏距离系数可以初步判断各组人群的相互关系及差异；根据聚类分析树枝图推出的人群间的相互关系受作者主观意识的影响，可信的结论应建立在多种聚类方法产生的结果一致的基础上；主成分分析的结果与选取的变量有一定关系，选取不同的变量组，其结果会受到影响。同聚类分析方法相比，主成分分析方法相对较好地反映了人群间的相互关系。本文研究结果提示，应慎重对待多变量统计方法得出的人群间相互关系的结论。

关键词：聚类分析；主成分分析；欧氏距离系数；头骨；测量性状

中图法分类号：Q983.3　　**文献标识码**：A　　**文章编号**：1000-3193（2007）04-0361-11

人类学研究中常用的多变量统计方法为“聚类分析法”和“主成分分析法”。多变量统计方法数学原理简单，结果直观，可以综合地提供各变量的信息，从而克服了传统的单一数据提供信息不足的缺点，因而在人类学研究领域，特别是在探索各地区人群间相互关系的研究中得到了较多的应用[1—4]。

聚类分析方法的基本过程是根据实体或者变量之间的相似程度，把它们逐步聚合成为一类，以此来推断各变量之间的相互关系。该方法的原理是采用一定的系数作为地区间的相似系数，先将所有的n个变量看成不同的n类，然后将性质最近（距离最近）的两类合并为一类；再将这n－1类中找到最接近的两类加以合并，依次类推，直到所有的变量被合并为一类。主成分分析的方法用于数据的降维，提取数据的主要信息，对变量进行综合判断和分类，第一主成分有最大的方差，后续成分，其可解释的方差越来越少。主成分分析不仅能对实体群进行分类，而且能够揭示变量在实体中的作用，所以应用较多[5,6]。

多变量统计分析方法虽然有很多优点，但是也存在不足之处，很多学者对由此方法得出的人群间相互关系的结果持怀疑的态度[7,8]，主要原因是缺乏对多变量统计分析方法的可靠性的验证。例如在聚类分析中，可用的测距方法有多种，可用的聚类方式也有多种，如类间

收稿日期：2006-10-10；　**定稿日期**：2007-04-05

基金项目：中国科学院知识创新工程重要方向项目（kzcx2-yw-106）；国家重点基础研究发展规划项目（2006CB806400）；国家基础科学人才培养基金（J0630965）资助

作者简介：吴秀杰，女，中国科学院古脊椎动物与古人类研究所副研究员，博士，主要从事古人类学研究。
Email：wuxiujie@ivpp.ac.cn

平均连锁法(Between-Groups)、类内平均连锁法(Within-Groups)、最近距离法(Nearest neighbor)、最远距离法(Furthest neighbor)、重心法(Centroid clustering)、中间距离法(Median clustering)、差离平方和法(Ward's Method)等等,那么得出的聚类树枝分析图就有多种,每种聚类图形的聚合水平分又分为多级,所以得出的数据的分组情况就会有更多种。又如,主成分分析结果与变量的选取有很大关系。变量组改变,主成分因子的负荷及贡献率会有很大不同。因此,有时需要进行因子分析,舍弃部分信息,提取主要因子,通过旋转等方式集中反映人群主要特征[5,6]。

理论上讲,如果所选的性状是由遗传因素决定的,那么无论采用何种研究方法,得出的人群间的相互关系的结果应该是一样的。如果结果不一致,就有以下几种可能:一种可能是所用的研究方法有问题,比如很多人在进行聚类分析时,并没有指明文中的聚类方法及聚合水平以及对不同聚类方法得出的结果的对比;第二种可能与样本的数量有关,如果样本量不足,其均数将不成正态分布;第三种可能与样本的变量有关,如在人类学研究中,头骨的人群差异比头后骨大得多;还有一种可能就是测量方法不规范,或者与测量者之间的误差有关。

有鉴于此,我们以生活在不同地区的9组人群的成年男性头骨为研究材料,由作者亲自测量,排除样本量不足、变量差异、测量方法不规范等因素所造成的干扰,探讨聚类分析和主成分分析在人类学研究中的价值。

1 材料和方法

1.1 研究材料

本文数据分析使用的标本为全新世不同地区、不同时代的9组成年男性头骨,标本总数为668例,标本来源于中科院古脊椎所和吉林大学边疆考古研究中心(见表1)。其中,长江以北的中原地区标本有河北组、陕西组、山西组和华北组,边疆地区标本有内蒙组、辽宁组和新疆组,长江以南的标本有云南组,国外标本有欧洲组。在这些标本中,河北组、陕西组、山西组、内蒙组、辽宁组、华北组和云南组属于蒙古人种类型,新疆组为蒙古人种和欧罗巴人种的混合类型,欧洲组属于欧罗巴人种类型[9]。

表1 本文使用的头骨材料(成年男性)

The cranial materials used in the paper (adult male)

组别	时代	遗址分布	标本例数
河北	新石器	阳原姜家梁	50
内蒙	青铜铁器	和林格尔土城子	86
辽宁	青铜铁器	北票喇嘛洞	58
陕西	青铜铁器	神木寨峁、陇县、铜川瓦窑堡	43
山西	青铜铁器	沂州游邀、大同	27
新疆	青铜铁器	鄯善洋海、尼勒克	57
华北	现代	华北地区	134
云南	现代	云南省境内	182
欧洲	近代	奥地利、捷克斯洛伐克	31
		合计	668

1.2 研究方法

参照近年来在全新世人群关系研究中常用的头骨性状[10—12],选取了14项测量性状进行多变量分析,包括颅长、颅宽、颅高、颅底长、面底长、上面高、面宽、鼻宽、鼻高、眶宽、眶高、额矢状弦长、顶矢状弦长和枕矢状弦长。对9组人群的头骨一一进行测量,分别计算出各地区人群的14项测量性状的平均值(见表2)。用SPSS软件对表2中的数据进行分析。

计算9组人群的欧氏距离系数和绝对值距离系数,比较采用不同的距离系数得出的人群间相互关系的差异。聚类分析,得出7种聚类方式(类间平均连锁法、类内平均连锁法、最

近距离法、最远距离法、重心法、中间距离法、差离平方和法)的聚合树枝图，比较不同聚合水平和不同聚类方法人群的分组情况是否具有一致性。

为验证主成分分析方法与所选变量的关系，本文设计了4组不同的变量组合方法进行主成分分析：方法一包括所有的14项人类学研究中常用的头骨测量性状(见表2)；方法二选取表2中的头骨主要测量指标包括颅长、颅宽、颅高、颅底长、面底长、上面高、面宽、鼻宽、鼻高、眶宽和眶高11项；方法三选取表2中的头骨大体形态测量指标包括颅长、颅宽、颅高、颅底长、面底长、上面高和面宽7项；方法四选取表2中的头盖部6项测量指标包括颅长、颅宽、颅高、额矢状弦长、顶矢状弦长和枕矢状弦长。分别对四组变量组合的数据进行主成分分析，提取第一、第二主成分因子，根据各因子的负荷矩阵计算出的各组人群的因子得分，绘制出二维分布图，比较四组变量组合得出的主成分分析的结果是否一致。

表2　本文使用的9组人群颅骨测量数值

The cranial metric data of nine populations　　(mm)

项目	内蒙	辽宁	新疆	陕西	山西	河北	华北	云南	欧洲
颅长	177.6	178.9	182.4	182.4	180.5	176.9	176.9	176.8	174.2
颅宽	143.1	143.9	139.6	140.6	144.0	135.2	137.1	137.6	148.2
颅高	141.7	141.6	135.4	138.6	138.9	137.5	136.3	132.0	131.4
颅底长	102.0	105.0	101.7	100.5	102.1	103.3	98.7	95.8	98.6
面底长	97.5	98.0	99.6	102.4	97.0	99.9	94.0	91.4	93.2
上面高	70.2	71.9	67.5	69.8	72.2	71.8	72.2	69.0	66.7
面宽	136.4	136.8	134.3	134.7	137.8	135.5	132.6	130.8	131.8
鼻宽	27.1	26.5	25.8	26.7	27.4	26.9	25.2	26.0	25.2
鼻高	53.3	55.8	51.3	53.9	54.2	54.7	55.3	53.4	49.7
眶宽	42.9	42.7	42.9	41.8	42.9	43.4	40.6	40.6	41.9
眶高	33.4	35.1	31.9	34.0	34.2	33.9	35.4	35.4	32.5
额矢状弦长	113.6	112.5	110.5	113.0	113.4	111.0	110.8	109.3	107.8
顶矢状弦长	111.6	120.7	115.9	115.4	114.5	111.6	112.1	110.7	109.3
枕矢状弦长	98.0	103.0	96.8	99.5	99.8	95.2	97.5	96.1	91.4

2　结　果

2.1　距离系数

2.1.1　欧氏距离

采用欧氏距离(Euclidean distance)作为地区间的相似系数，对表2中的数据进行聚类分析，得出的9组人群间的相互关系为(见表3)：内蒙、山西、陕西、河北和辽宁之间的距离较小，表明它们之间关系较为密切；华北和云南的距离较小，而与欧洲之间的距离较大。新疆与欧洲的距离5.531，与内蒙之间的系数为4.327，与陕西之间的系数为3.503，新疆与陕西人群的关系较其它人群密切。

表3　欧氏距离系数

Euclidean distance coefficients

	内蒙	辽宁	新疆	陕西	山西	河北	华北	云南
辽宁	3.975	0.000						
新疆	4.327	5.671	0.000					
陕西	3.164	3.867	3.503	0.000				
山西	2.188	3.013	4.814	2.982	0.000			
河北	3.145	4.721	4.460	3.855	3.620	0.000		
华北	5.143	5.602	5.719	4.725	5.325	4.611	0.000	
云南	6.093	7.335	5.877	5.742	6.494	5.717	2.968	0.000
欧洲	6.634	8.719	5.531	7.257	7.682	6.886	6.249	4.974

2.1.2　绝对值距离

采用绝对值距离(City block distance)作为地区间的相似系数，得出的9组人群间的相互关系(见表4)的结果与欧氏距离系数基本相似。

2.2 聚类分析

2.2.1 类间平均连锁法

表 4 绝对值距离系数

City block distance coefficients

	内蒙	辽宁	新疆	陕西	山西	河北	华北	云南
辽宁	11.209							
新疆	14.200	19.060						
陕西	10.214	13.464	11.072					
山西	6.720	9.307	15.448	8.949				
河北	9.633	14.165	14.054	12.310	11.282			
华北	16.856	18.273	17.794	16.063	17.136	12.814		
云南	19.407	24.711	18.754	19.033	22.726	17.630	8.960	
欧洲	22.841	30.614	18.418	24.949	27.123	24.091	18.931	15.826

图 1A 为根据欧氏距离系数，采用类间平均连锁法对表 2 中的数据进行聚类分析得出的树枝图。当聚合水平 > 5 时，9 组人群分成六组：第一组为内蒙、山西和陕西；第二组为河北；第三组为辽宁；第四组为新疆；第五组为华北和云南；第六组为欧洲。当聚合水平 > 10 时，9 组人群分成四组：第一组为内蒙、山西、陕西、河北和辽宁；第二组为新疆；第三组为华北和云南；第四组为欧洲。当聚合水平 > 15 时，9 组人群分成三组：第一组为内蒙、山西、陕西、河北、辽宁和新疆；第二组为华北和云南；第三组为欧洲。当聚合水平 > 20 时，9 组人群分成二组：欧洲地区人群为一组，其它 8 组人群聚合成一组。

2.2.2 类内平均连锁法

图 1B 为根据欧氏距离系数，采用类内平均连锁法对表 2 中的数据进行聚类分析得出的树枝图。当聚合水平 > 5 时，9 组人群分成五组：第一组为内蒙、山西、陕西和河北；第二组为辽宁；第三组为新疆；第四组为华北和云南；第五组为欧洲。当聚合水平 > 10 时，9 组人群分成三组：第一组为内蒙、山西、陕西、河北、辽宁和新疆；第二组为华北和云南；第三组为欧洲。当聚合水平 > 15 时，9 组人群分成三组：第一组为内蒙、山西、陕西、河北、辽宁和新疆；第二组为华北和云南；第三组为欧洲。当聚合水平 > 20 时，9 组人群分成二组：内蒙、山西、陕西、河北、辽宁和新疆聚合为一组；华北、云南和欧洲聚合为一组。

2.2.3 最近距离法

图 1C 为根据欧氏距离系数，采用最近距离法对表 2 中的数据进行聚类分析得出的树枝图。当聚合水平 > 5 时，9 组人群分成六组：第一组为内蒙、山西和陕西；第二组为辽宁；第三组为河北；第四组为新疆；第五组为华北和云南；第六组为欧洲。当聚合水平 > 10 时，9 组人群分成三组：第一组为内蒙、山西、陕西、河北和新疆；第二组为华北和云南；第三组为欧洲。当聚合水平 > 20 时，9 组人群分成二组：欧洲聚合为一组；其它 8 组人群聚合成一组。

2.2.4 最远距离法

图 1D 为根据欧氏距离系数，采用最远距离法对表 2 中的数据进行聚类分析得出的树枝图。当聚合水平 > 5 时，9 组人群分成五组：第一组为内蒙、山西和河北；第二组为辽宁；第三组为华北和云南；第四组为新疆和陕西；第五组为欧洲。当聚合水平 > 10 时，9 组人群分成四组：第一组为内蒙、山西、河北和辽宁；第二组为华北和云南；第三组为新疆和陕西；第四组为欧洲。当聚合水平 > 15 时，9 组人群分成三组：第一组为内蒙、山西、河北和辽宁；第二组为华北、云南、新疆和陕西；第三组为欧洲。当聚合水平 > 20 时，9 组人群分成二组：欧洲聚合为一组；其它 8 组人群聚合成一组。

2.2.5 重心法

图 1E 为根据欧氏距离系数，采用重心法对表 2 中的数据进行聚类分析得出的树枝图。当聚合水平 > 5 时，9 组人群分成四组：第一组为内蒙、山西、陕西、河北和辽宁；第二组为华

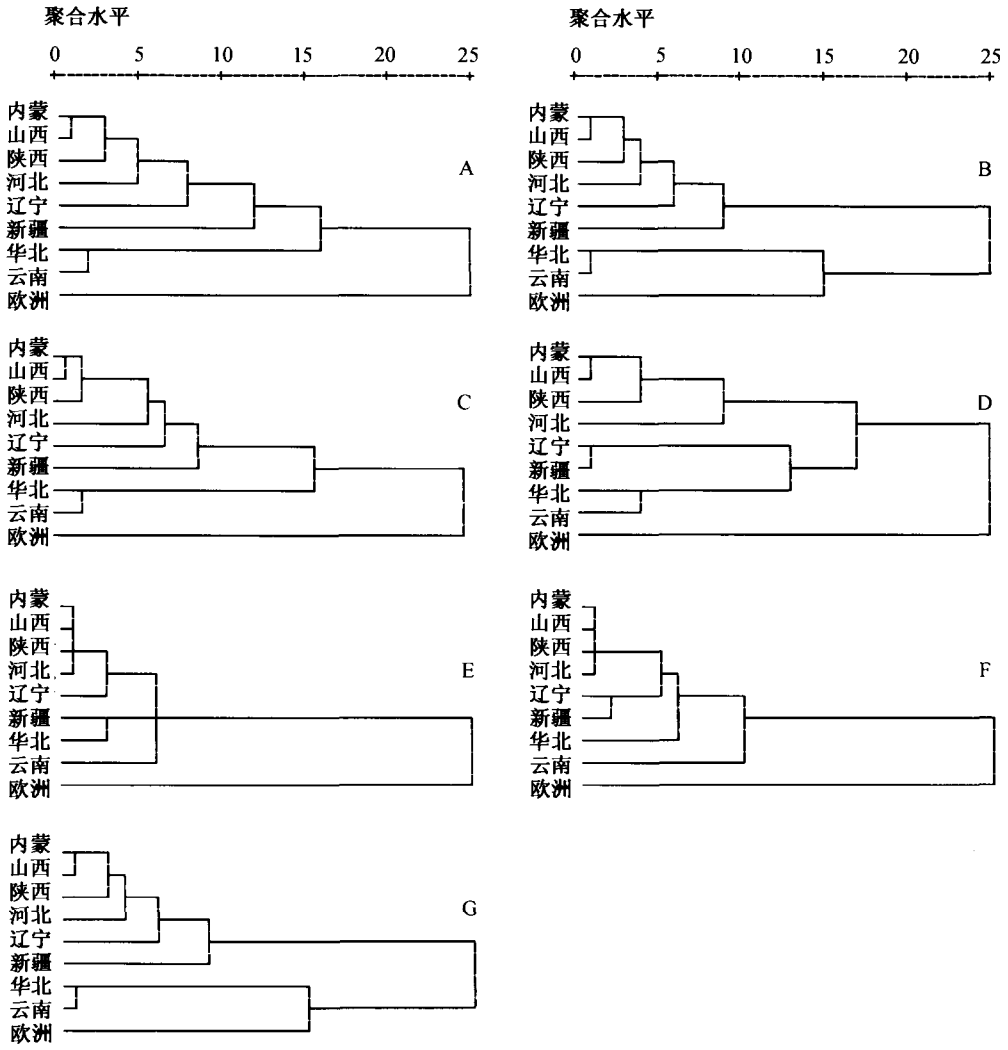

图 1　采用不同的聚类方法得出的九组人群测量数据的聚类树枝图

Dendragram drew of metric data of nine populations using different cluster analysis methods

A.类间平均连锁法（Between-Groups）；B.类内平均连锁法（Within-Groups）；C.最近距离法（Nearest neighbor）；D.最远距离法（Furthest neighbor）；E.重心法（Centroid clustering）；F.中间距离法（Median clustering）；G.差离平方和法（Ward's Method）

北和云南；第三组为新疆；第四组为欧洲。当聚合水平 > 10 时，9 组人群分成二组：欧洲聚合为一组；其它 8 组人群聚合成一组。

2.2.6　中间距离法

图 1F 为根据欧氏距离系数，采用中间距离法对表 2 中的数据进行聚类分析得出的树枝图。当聚合水平 > 5 时，9 组人群分成五组：第一组为内蒙、山西、陕西和河北；第二组为华北和云南；第三组为新疆；第四组为辽宁；第五组为欧洲。当聚合水平 > 10 时，9 组人群分成三组：第一组为内蒙、山西、陕西、河北、华北、云南和新疆；第二组为辽宁；第三组为欧洲。当聚合水平 > 15 时，9 组人群分成二组：欧洲聚合为一组；其它 8 组人群聚合成一组。

2.2.7 差离平方和法

图 1G 为根据欧氏距离系数,采用差离平方和法对表 2 中的数据进行聚类分析得出的树枝图。当聚合水平 > 5 时,9 组人群分成五组:第一组为内蒙、山西、陕西和河北;第二组为辽宁;第三组为新疆;第四组为华北和云南;第五组为欧洲。当聚合水平 > 10 时,9 组人群分成三组:第一组为内蒙、山西、陕西、河北、辽宁和新疆;第二组为华北和云南;第三组为欧洲。当聚合水平 > 20 时,9 组人群分成二组:内蒙、山西、陕西、河北、辽宁和新疆人群聚合成一组;华北、云南和欧洲人群聚合为一组。

2.3 主成分分析

2.3.1 14 项测量性状主成分分析

这项主成分分析使用了表 2 中的所有 14 项头骨测量数据。从表 5 罗列的主成分因子负荷及贡献率来看,第一主成分和第二主成分对变量信息的贡献率分别为 52.4% 和 21.6%。其中颅底长、面宽和眶宽在第一主成分具有较大的因子载荷,说明分布在图 2A 中右侧的辽宁、山西、陕西、河北、内蒙和新疆组有较大的颅底长和面宽。眶高、鼻高和上面高在第二主成分具有较大的因子载荷,说明分布在图 2A 中上部的华北、云南和辽宁组有较大的眶高和鼻高。从图 2A 根据第一和第二主成分因子得分绘制的各组人群相互关系图可以看出,内蒙、陕西、山西、河北和辽宁组距离较近;华北和云南组较近;新疆和欧洲组远离其它各组。

表 5 主成分分析因子负荷及贡献率

Principal components analysis (PCA) loadings

	14 项测量性状		11 项测量性状		7 项测量性状		6 项测量性状	
	PCA 1	PCA 2	PCA 1	PCA 2	PCA 1	PCA 2	PCA 1	PCA 2
贡献率	52.4%	21.6%	49.1%	26.4%	56.1%	17.9%	62.3%	17.5%
颅长	0.103	-0.047	0.119	-0.041	0.154	-0.028	0.197	-0.252
颅宽	0.051	-0.164	0.049	-0.185	0.112	0.746	0.009	0.934
颅高	0.118	0.049	0.151	0.087	0.232	-0.035	0.236	0.114
颅底长	0.143	-0.065	0.180	-0.025	0.248	0.106		
面底长	0.137	-0.096	0.168	-0.066	0.206	-0.015		
上面高	0.021	0.224	0.044	0.267	0.085	-0.452		
面宽	0.145	-0.034	0.183	0.007	0.258	0.099		
鼻宽	0.111	0.004	0.146	0.044				
鼻高	0.007	0.258	0.024	0.295				
眶宽	0.149	-0.177	0.185	-0.143				
眶高	-0.081	0.308	-.088	0.319				
额矢状弦长	0.112	0.059					0.241	0.014
顶矢状弦长	0.097	0.025					0.226	0.071
枕矢状弦长	0.076	0.133					0.253	-0.005

2.3.2 11 项测量性状主成分分析

这项主成分分析使用了颅长、颅宽、颅高、颅底长、面底长、上面高、面宽、鼻宽、鼻高、眶宽、眶高 11 项头骨测量数据。从表 5 罗列的主成分因子负荷及贡献率来看,第一主成分和第二主成分对变量信息的贡献率分别为 49.1% 和 26.4%。其中眶宽、面宽和颅底长在第一主成分具有较大的因子载荷,说明分布在图 2B 中右侧的辽宁、山西、陕西、河北、内蒙和新疆组有较大的眶宽和面宽。眶高、鼻高和上面高在第二主成分具有较大的因子载荷,说明分布在图 2B 中上部的北、云南和辽宁组有较大的眶高和鼻高。从图 2B 根据第一和第二主成分因子得分绘制的各组人群相互关系图可以看出,内蒙、陕西、山西、河北和辽宁组距离较近;华北和云南组较近;新疆和欧洲组远离其它各组。

2.3.3 7 项测量性状主成分分析

这项主成分分析使用了颅长、颅宽、颅高、颅底长、面底长、上面高和面宽 7 项头骨测量

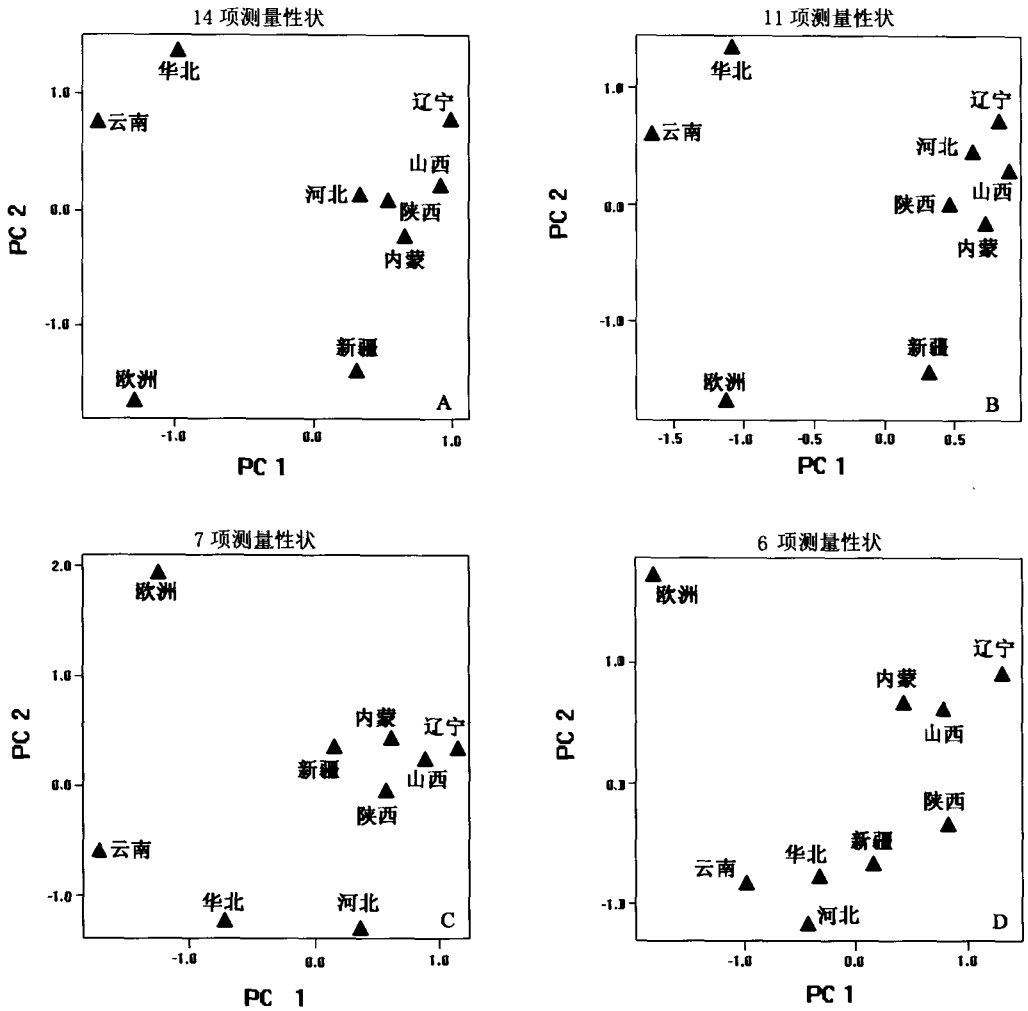

图 2　九组人群头骨测量数据四种主成分分析方法两维坐标分布图比较

Two dimensional distributions of four PCA scores methods from skull metrical data of nine populations

数据。从表 5 罗列的主成分因子负荷及贡献率来看，第一主成分和第二主成分对变量信息的贡献率分别为 56.1% 和 17.9%。其中面宽和颅底长在第一主成分具有较大的因子载荷，说明分布在图 2C 中右侧的辽宁、山西、内蒙、陕西、新疆和河北组有较大的面宽和颅底长。颅宽和上面高在第二主成分具有较大的因子载荷，说明分布在图 2C 中上部的欧洲组有较大的颅宽和上面高。从图 2C 根据第一和第二主成分因子得分绘制的各组人群相互关系图可以看出，辽宁、山西、内蒙、陕西和新疆组距离较近；华北位于云南和河北之间；欧洲组远离其它各组。

2.3.4　6 项测量性状主成分分析

这项主成分分析使用了颅长、颅宽、颅高、额矢状弦长、顶矢状弦长和枕矢状弦长 6 项头骨测量数据。从表 5 罗列的主成分因子负荷及贡献率来看，第一主成分和第二主成分对变量信息的贡献率分别为 62.3% 和 17.5%。其中枕矢状弦长和额矢状弦长在第一主成分具

有较大的因子载荷，说明分布在图 2D 中右侧的辽宁、山西、陕西和内蒙组有较大的枕矢状弦长和额矢状弦长。颅宽和颅高在第二主成分具有较大的因子载荷，说明分布在图 2D 中上部的欧洲组有较大的颅宽和颅高。从图 2D 根据第一和第二主成分因子得分绘制的各组人群相互关系图可以看出，辽宁、山西和内蒙组关系较近；陕西、新疆、华北、云南和河北组关系较近；欧洲组远离其它各组。

3 分析和讨论

在研究现代人群相互关系时，主要是依据头骨、牙齿、或头后骨的一些标志性的形态特征，采用单性状分析、聚类分析或主成分分析的方法，对各变量的测量或非测量性状进行统计、比较和分析，以此来推断各地区人群的亲缘关系，其中应用范围最广泛的是头骨的测量性状。在应用多变量统计分析方法之前，首先假设所选用的头骨测量性状主要是由遗传因素决定的，如果所选性状受环境因素的影响较大，那么所得出的各地区人群之间相互关系的结果就会不可靠。一些学者认为头骨的测量性状是由遗传因素决定的，可以用来反映人群之间的生物学关系[13—14]。还有一些学者认为，测量性状是由遗传和环境因素共同作用引起的，基因的突变、漂移、选择、交流和环境的改变，对测量性状有一定程度的影响[15—17]。从目前研究来看，多数学者倾向于头骨的测量性状主要受遗传因素的影响，可以用来分析各地区人群之间的相互关系。

目前，许多学者在利用多变量研究方法进行各地区人群相互关系的分析时，并没有考虑到此种方法是否准确可靠，这在一定程度上影响了人们对其研究的认可。本文通过对生活在不同地理区域的 9 组人群头骨的各项测量数据的聚类分析和主成分分析，对多变量统计方法提出以下看法。

3.1 欧氏距离系数可以初步判断各地人群的相互关系

距离系数是反映两个实体相异和亲疏度量的指标，距离系数有多种，包括：欧氏距离、欧氏距离平方、绝对值距离、马氏距离等等[18]。本文作者曾实验了各种距离系数，得出的人群间的相互关系的结果基本相同。其中人类学研究中常用的是欧氏距离，欧氏距离系数不仅可以初步判断各组人群间的相互关系，而且还可以分析人群间的差异。欧氏系数越大，表明人群间的差异越大；系数越小，表明人群间的关系较近。本文欧氏距离系数分布显示：内蒙、山西、陕西、河北和辽宁之间的关系较为密切；华北和云南关系较为密切；欧洲远离其它各组；新疆与陕西、内蒙等组的关系近于欧洲。内蒙、山西、陕西、河北和辽宁组属于考古遗址出土的蒙古人种成员，华北和云南组属于现代蒙古人种成员，欧洲组属于欧罗巴人种成员，新疆组属于蒙古人种与欧罗巴人种杂交类型。从本文研究结果来看，由欧氏距离系数得出的人群关系比较符合人种学研究的成果[9,12]。

3.2 由聚类分析树枝图得出的人群关系含有很大的主观成分

聚类分析的结果依赖于相似系数和聚类方法的选取。选取不同的相似系数和使用不同的聚类方法可能会给出不同的树枝聚类图[6]。在本文研究种，当聚合水平 > 20 时，7 种聚类方式得到的树枝图结果有两种：一种结果把欧洲组单列为一组，其它 8 组人群聚合在一起；另外一种结果把内蒙、山西、陕西、辽宁和新疆聚合在一起，而把华北、云南和欧洲聚合在一起。当聚合水平 > 15 时，7 种聚类方式得到的树枝图结果有三种：第一种结果把内蒙、山西、

陕西、河北、辽宁和新疆聚合在一起，华北和云南聚合在一起，欧洲单列为一组；第二种结果把内蒙、山西、河北和辽宁聚合为一组，把华北、云南、新疆和陕西聚合为一组，欧洲单列为一组；第三种结果欧洲为一组，其它8组人群为一组。当聚合水平＞10时，7种聚类方式得到的树枝图结果有五种：第一种结果把内蒙、山西、陕西、河北和辽宁聚合为一组，新疆为一组，华北和云南为一组，欧洲为一组；第二种结果把内蒙、山西、陕西、河北、辽宁和新疆聚合为一组，华北和云南为一组，欧洲为一组；第三种结果把内蒙、山西、河北和辽宁聚合为一组，把华北和云南聚合为一组，把新疆和陕西聚合为一组，欧洲为一组；第四种结果把欧洲聚合为一组，其它8组人群聚合成一组；第五种结果把内蒙、山西、陕西、河北、华北、云南和新疆聚合为一组，辽宁为一组，欧洲为一组。当聚合水平＞5时，7种聚类方式得到的树枝图结果有四种：第一种把内蒙、山西和陕西聚合为一组，河北为一组，辽宁为一组，新疆为一组，华北和云南为一组，欧洲为一组；第二种结果把内蒙、山西、陕西和河北聚合为一组，辽宁为一组，新疆为一组，华北和云南为一组，欧洲为一组；第三种结果把内蒙、山西和河北聚合为一组，辽宁为一组，华北和云南为一组，新疆和陕西聚合为一组，欧洲为一组；第四种结果把内蒙、山西、陕西、河北和辽宁聚合为一组，华北和云南聚合为一组，新疆为一组，欧洲为一组。

在以上各种树枝图中，与欧氏距离系数最为接近的结果有三项：第一项，以组间法进行聚类分析，当聚合水平＞10时；第二项，以重心法行聚类分析，当聚合水平＞5时；第三项，以最远距离法聚类分析，当聚合水平＞10时。前两项结果把9组人群分为四组：内蒙、山西、陕西、河北和辽宁为一组；华北和云南为一组；新疆为一组；欧洲为一组。第三项结果把内蒙、山西、河北和辽宁聚合为一组，华北和云南聚合为一组，新疆和陕西聚合为一组，欧洲单列为一组。

在得到数枝图后，对于聚合水平的选取以及人群的分组情况，目前还没有分类的统计学依据和被普遍接受的共识，很多研究者在对聚类图形进行分析时，只是选择其中的适合研究者讨论分析的结果，具有很大的主观性。正确的方法应对比不同的聚类方法得出的结果，只有在各种聚类方法得出的树枝图基本一致的基础上，其结论才可信。

3.3　主成分分析方法与所选变量有很大关系

主成分分析的方法可以提取数据的主要信息，对其进行分类。在本文使用的4组不同变量的组合中，变量不同，其提取的具有较大因子载荷的头骨测量项目不同，各地人群之间的相互关系亦有所改变。其中，使用14项和11项头骨测量数据得到的各地人群的相互关系比较接近，二者都把内蒙、山西、陕西、河北和辽宁聚合为一组，华北和云南比较接近，新疆为一组，欧洲为一组，此项结果与欧氏距离系数得出的结果一致。使用7项和6项头骨测量数据得出的各地人群间的相互关系基本与客观事实符合，除了把河北组单列为一组以外，这可能与河北组属于新石器时代人群有关。可见，主成分分析方法与所选的变量有一定关系，但同聚类分析比较，主成分分析方法相对较好地反映人群之间的相互关系。目前越来越多的学者倾向使用主成分分析方法进行人类学研究[19—21]，通过对比不同的变量组合，探讨各地人群的亲缘关系。

综上所述，聚类分析和主成分分析是探索性的分析方法，其结论受研究者主观意识的影响。在进行聚类分析研究时，需要对比不同的聚类方法和不同的聚合水平下的树枝图；在进行主成分分析时，需要对比不同的变量组合。只有在各种研究方法一致的基础上，得出的结论才可靠。

致谢：此文为本文第一作者博士论文的一部分，感谢吴新智院士、刘武研究员和张银运研究员的悉心指导和帮助，另外也非常感谢吉林大学边疆考古研究中心提供的部分研究材料。

参考文献：

[1] 王令红. 中国新石器时代和现代居民的时代变化和地理变异－颅骨测量性状的统计分析研究[J]. 人类学学报，1986，5：243-258.

[2] 张振标. 现代中国人体质特征及其类型的分析[J]. 人类学学报，1988，7：314-323.

[3] 刘武，铃木基治. 亚洲地区人类群体亲缘关系－活体测量数据统计分析[J]. 人类学学报，1994，13：265-279.

[4] Kidder JH，Durband AC. A reevaluation of the metric diversity within *Homo erectus*[J]. J Hum Evol，2004，46：297-313.

[5] 张文彤. SPPS 11 统计分析教程[M]. 北京：希望电子出版社，2002：111-113.

[6] 陈铁梅. 定量考古学[M]. 北京：北京大学出版社，2005：1-287.

[7] Rightmare GP. Cranial measurements and discrete traits compared in distance studies of African Negro skulls[J]. Hum Biol，1972：263-276.

[8] Guglielmino-Matessi CR，Gluckman P，Cavalli-Sforza LL. Climate and the evolution of skull metrics in man[J]. Am J Phys Anthropol，1979，50：549-564.

[9] 韩康信，潘其风. 中国古代人种成分研究[J]. 考古学报，1984，2：245-263.

[10] 吴汝康，吴新智，张振标. 人体测量方法[M]. 北京：科学出版社，1984：1-172.

[11] 张振标. 中国新石器时代人类遗骸[C]. 见：吴汝康，吴新智，张森水主编. 中国远古人类. 北京：科学出版社，1989：62-80.

[12] 韩康信. 中国夏、商、周时期人骨种族特征之研究[C]. 见：中国社会科学院考古研究所编著. 新世纪的中国考古学. 北京：科学出版社，2005：925-966.

[13] Howells WW. Cranial Variation in Man：A Study by Multivariate Analysis of Patterns of Difference among Recent Human Populations[M]. Cambridge：Harvard University Press，1973：1-259.

[14] Devor EJ. Transmission of human craniofacial dimensions[J]. J Craniofac Genet Dev Biol，1987，7：95-96.

[15] Carlson DS. Temporal variation in Prehistoric Nubian crania[J]. Am J Phys Anthropol，1977. 45：467-484.

[16] Van VG，Schaafoma W. Advances in the quantitative analysis of skeletal morphology[C]. In：Sanders S，Katzenberg M，eds. Skeletal biology of past peoples：research methods. New York：Wiley and Liss，1992：225-257.

[17] Hemphill B. Biological affinities and adaptations of Bronze Age Bactrians：IV. A radiometric investigation of Bactrian origins[J]. Am J Phys Anthropol，1999，108：173-192.

[18] 李春喜，王志和，王文林. 生物统计学[M]. 北京：科学出版社，2000：229-246.

[19] 刘武，张银运. 中国直立人形态特征的变异-颅骨测量数据的分析[J]. 人类学学报，2005，24：121-136.

[20] Neves WA，Hubbe M. Cranial morphology of early Americans from Lagoa Santa，Brazil：Implications for the settlement of the New World[J]. P Natl Acd Sci，2005，102：18309-18314.

[21] Wu Xiujie，Schepartz L，Falk D *et al*. Endocast of Hexian *Homo erectus* from south China[J]. Am J Phys Anthropol，2006，130：445-454.

An Examination of Cluster and Principle Component Analysis on the Study of Anthropology

WU Xiu-jie[1], ZHANG Quan-chao[1,2], LI Hai-jun[1,3]

(1. *Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing* 100044;
2. *Research Center for Chinese Frontier Archaeology of Jilin University, Changchun* 13001;
3. *Graduate school of Chinese Academy of Sciences, Beijing* 100049)

Abstract: The Multivariate analysis can synthesize the database and supply the direct information, so more and more anthropologists prefer the method to analyze the relationship among the different populations. Because few people tested the method, some researchers still suspected the result from the Multivariate analysis. In order to conduct Multivariate analysis on the study of anthropology, we chose adult male skulls (n = 668) of nine populations related to the different areas. These populations included: Hebei, Inner Mongolia, Liaoning, Shaanxi, Shanxi, Xinjiang, Huabei, Yunnan and Europe. Fourteen standard linear measures were culled to do cluster and principal components analysis. The relationship and difference of the populations are very similar comparison the result from Euclidean distance coefficient and City block distance. The primary results of this study indicate that Euclidean distance coefficient is useful for primarily judging the relationship and difference of the populations. The dendragrams drew of metric data of nine populations using different cluster analysis methods were varied. It is uncertain to determine the relationship of the populations only according to the cluster dendrogram, except the results from all kinds of cluster methods are consistent. With four PCA scores methods from skull metrical data, the distributions of nine populations did not change a lot. The principal components analysis is associated with the variables. When the variables change, the component matrix and the total variance loadings change too. Compared with cluster analysis, principal components analysis is better to explain the relationship of the populations. It suggests that the conclusion from multi-variables analysis should be considered carefully.

Key words: Cluster analysis; Principle component analysis; Euclidean distance coefficient; Skull; Metric traits