第46卷 第2期 气 象 学 报 Vol. 46, No. 2
1988 年 5 月 ACTA METEOROLOGICA SINICA May, 1988

气候分析应用聚类方法的评价*

谭 冠 日**

(中山大学大气科学系)

提 要

对聚类分析的华德法、重心法和平均联系法的理论分析指出:在组间距离相同的条件下,华德法优先合并小的组,最终产生大小较为均匀的组,这不能准确地反映气候的现实。在组内平方和增量相同的条件下,重心法优先合并大的组,最终产生一个很大的组和很多极小的组,背离气候的现实。平均联系法基本上避免了上述偏歧,产生凝聚、方差均匀的组,较好地反映气候现实。中国东部大陆夏季降水的分区证实了以上结论。

一、引 言

聚类分析技术以数值方法将纪录进行分类,在日益走向客观化、定量化的气候分析中得到越来越广泛的应用。聚类分析的某些环节又离不开人的判断,既可引入分析人员的经验,又可能掺入主观成份[1,2]。尽管数学界已有人注意到不同聚类方法的偏歧[3,4],但没有把各种方法进行对比评价。气象学者在采用某种聚类分析方法时,一般并未注意到它对研究结果可能带来的影响,甚至造成的假象。

本文评价阶梯(hierachical)和聚集(agglomerative)型聚类的华德法(Ward's)、重心法(centroid)和平均联系法(average linkage),因为它们代表了几种不同类型,并在气候分析中采用较多[2,5-8]。电脑的统计软件包往往提供它们的子程序,便于应用。我们先借统计理论考察它们的性质、相互联系和偏歧。

二、理 论 分 析

1. 华德法[9]的原则是每合并一组都使当步的组内平方和

$$W=\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{N_k}(X_{ijk}-X_{\cdot jk})^2 \tag{1}$$

达到极小。X_{ijk}是在K个组中的第k组、J个变量中的第j个变量、N_k个纪录中的第i个纪录的数值。下标中一点代表对该下标求平均。后文一般略去求和号的上下限,除非必要时才写。

* 本文于1986年5月29日收到,1986年9月3日收到修改稿。
** 此文系作者在美国特拉华大学气候研究中心工作时完成的。

(1) 式右端加、减一项 $X_{\cdot j\cdot}$(合并后的组第 j 个变量的重心),化简后得

$$W=\sum_{k}\sum_{j}\sum_{i}(X_{ijk}-X_{\cdot j\cdot})^2-\sum_{k}N_k\sum_{j}(X_{\cdot j\cdot}-X_{\cdot jk})^2 \tag{2}$$

假定 1 组含 N_1 个纪录,将与 2 组(含 N_2 个纪录)合并。并组前这两组的组内平方和是

$$W=\sum_{k=1}^{2}\sum_{j=1}^{J}\sum_{i=1}^{N_k}(X_{ijk}-X_{\cdot j\cdot})^2-\sum_{k=1}^{2}N_k\sum_{j=1}^{J}(X_{\cdot j\cdot}-X_{\cdot jk})^2$$

$$=\sum_{j=1}^{J}\sum_{i=1}^{N_1}(X_{ij1}-X_{\cdot j\cdot})^2+\sum_{j=1}^{J}\sum_{i=1}^{N_2}(X_{ij2}-X_{\cdot j\cdot})^2-\sum_{k=1}^{2}N_k\sum_{j=1}^{J}(X_{\cdot j\cdot}-X_{\cdot jk})^2 \tag{3}$$

并组后的组内平方和是

$$W'=\sum_{j=1}^{J}\sum_{i=1}^{N_1}(X_{ij1}-X_{\cdot j\cdot})^2+\sum_{j=1}^{J}\sum_{i=1}^{N_2}(X_{ij2}-X_{\cdot j\cdot})^2 \tag{4}$$

此式右端两项分别包含原属 1,2 组的纪录。

华德法并组的判据是使W的增量 $\Delta W=W'-W$ 极小,而 ΔW 只与行将合并的两组有关。利用(3)、(4)式,得

$$\Delta W=N_1\sum_{j=1}^{J}(X_{\cdot j\cdot}-X_{\cdot j1})^2+N_2\sum_{j=1}^{J}(X_{\cdot j\cdot}-X_{\cdot j2})^2 \tag{5}$$

新组(称为 3 组)的重心 $X_{\cdot j\cdot}$ 与原来两组的重心 $X_{\cdot j1}, X_{\cdot j2}$ 的距离平方是

$$D_{13}^2=\sum_{j=1}^{J}(X_{\cdot j1}-X_{\cdot j\cdot})^2 \tag{6}$$

$$D_{23}^2=\sum_{j=1}^{J}(X_{\cdot j2}-X_{\cdot j\cdot})^2 \tag{7}$$

利用(6)、(7)式,(5)式又可以写成

$$\Delta W_{12}=N_1D_{13}^2+N_2D_{23}^2 \tag{8}$$

(8) 式表明原来组与新组重心的距离乃是华德法并组的一种根据。 Anderberg[3]曾从另一途径推导出

$$\Delta W_{12}=\frac{N_1N_2}{N_1+N_2}D_{12}^2 \tag{9}$$

这式表明原来两组的距离也是华德法并组的一种根据。

(8) 式更表明,纪录数目对 ΔW_{12} 发生影响。(9) 式也给出同样结论。比如,从一个特定的组到两个组的距离相等,这个组将与小的组合并。因为华德法追求组内平方和极小,组越小,一般来说组内平方和也小。在其他条件(距离)相同的条件下,优先合并小的组。小组增长快,大组往往维持原状,最终形成各组的大小相差不多。这是一种偏岐。有些文献[4,10]已注意到这一现象。(9) 式及本文所推导的 (8) 式对此作出了理论上的证明。采用华德法进行气候分区,使站点稠密地方的气候区较小,站点稀疏的地方气候区较

大，这往往不能准确地反映气候的现实。

2．重心法的原则是合并距离最近的组。距离的定义为(6)或(7)式。此法看来有理，事实上有些缺点十分严重。由(9) 式可知

$$\Delta W_{13}=\frac{N_1N_3}{N_1+N_3}D^2_{13} \tag{10}$$

如果1组与3组合并后，组内平方和增量ΔW_{13}和1组与2组合并后的增量ΔW_{12}相等，按华德法，1组与哪个组合并并无定论，但重心法却使1组与较大的组合并，因为从(9)、(10)两式可导出

$$\frac{D^2_{12}}{D^2_{13}}=\frac{\frac{N_1}{N_2}+1}{\frac{N_1}{N_3}+1} \tag{11}$$

若$N_2>N_3$，则$D_{12}<D_{13}$，1组将与2组合并。这证明了在其他条件（平方和的增量ΔW）相同的条件下，重心法合并大的组。大组容易再增大，象滚雪球一样，小组往往维持原状，最终形成一个很大的组和很多极小的组。这是一种严重的偏歧。采用重心法进行气候分区，将会得到一个很大（包含很多站点）的区，和很多极小（往往只有一两个站点）的区。

此外，重心法还有“链式反应”：在并组的过程中把新组的重心向某一定方向拖曳[10,11]；在并组过程中，重心间的距离并不总是一步比一步大，有时反而减少，存在着微小波动[12]；占用机时特别长。

作者曾用重心法作为一种参考、对照的方法进行美国南方气团分类，结果与气候现实大相径庭[13]。

3．平均联系法的指标是一个组内所有纪录对另一组内所有纪录的欧氏距离平方的平均值

$$L_{12}=\frac{1}{N_1N_2}\sum_{p=1}^{N_1}\sum_{q=1}^{N_2}D^2_{pq} \tag{12}$$

下标p，q分别表示1组、2组内纪录的序号。

对这个指标的分析可揭示此法的性质及它与其他两法的异同。

1，2组的组内平方和可分别写成

$$W_1=\sum_p\sum_j X^2{}_{pj1}-\frac{1}{N_1}\sum_j(\sum_p X_{pj1})^2 \tag{13}$$

$$W_2=\sum_q\sum_j X^2{}_{qj2}-\frac{1}{N_2}\sum_j(\sum_q X_{qj2})^2 \tag{14}$$

两组间的距离可展开为

$$D^2_{12}=\frac{1}{N_1^2}\sum_j(\sum_p X_{pj1})^2-\frac{2}{N_1N_2}\sum_j(\sum_p X_{pj1}\sum_q X_{qj2})+\frac{1}{N_2^2}\sum_j(\sum_q X_{qj2})^2 \tag{15}$$

(12)式展开成

$$L_{12}=\frac{1}{N_1N_2}\sum_p\sum_q\sum_j(X_{pj1}-X_{qj2})^2$$

$$=\frac{1}{N_1N_2}\left[N_1\sum_q\sum_j X_{qj2}^2+N_2\sum_p\sum_j X_{pj1}^2-2\sum_p\sum_q\sum_j X_{pj1}X_{qj2}\right] \tag{16}$$

将(13)—(15)式代入(16)式，最后得到

$$L_{12}=\frac{W_1}{N_1}+\frac{W_2}{N_2}+D_{12}^2$$

$$=S_1^2+S_2^2+D_{12}^2 \tag{17}$$

其中S_1^2，S_2^2分别是 1，2 组的方差。

可见，“平均联系”事实上是两组的方差及两组间距离之和。平均联系法包含了D_{12}^2，也受纪录数目的影响(如同重心法)，但被两项方差所冲谈。两项方差则几乎不受纪录数目的影响，这比华德法优越（华德法考虑的是平方和，而不是方差）。平均联系法的结果介乎重心法和华德法之间，即不倾向于产生大小十分悬殊的组，也不倾向于产生大小均匀的组。它的指标既然含有方差，L_{12}的极小化在某种程度上寻求组内方差极小，因而使组间方差极大[14-16]，这正是聚类分析的目的。在阐述平均联系法的良好性质的同时，不能不注意到另一方面。它优先合并方差小的组，方差大的组趋于维持，最终产生的各组的方差相差不大。这是否符合气候的现实及聚类的目的，要视分析的具体对象和要求而定。

三、实 例 检 验

我们以我国东部大陆上夏季降水年际变化的同步性分区来检验上述理论推断。资料是 110°E 以东、40°N 以南大陆上 50 个站(图 1 所示) 1951—1980 年夏季(6—8 月) 的降水距平百分率[17]。

聚类过程中如何截取适当的组数，并无通用标准，考虑的尺度不同就可截取不同的组数。我们采取 R^2 的陡降作为标准，截取 5 组(区) 和 10 组(区) 两个方案，作三种方法的比较(表 1)。R^2 是组间平方和与总的平方和之比[4]，并组开始之前，$R^2=1$；最后并成一组之后，$R^2=0$；并组过程中，R^2 趋于减小。R^2 在某一步陡降，意味着这一步把差异较大的组合并了，应保留这一步以前的组数。

无论是 5 区还是 10 区方案，重心法结果都很不合理，无须再作讨论。5 区方案华德法和平均联系法接近，只有一点差别(图 1)。

5 区方案平均联系法把济南划入华北一区，华德法则划入黄河下游及山东一区(简称南区)。我们从济南与上述两区各站的相关系数来检验划分的合理性(表 2)。

济南与华北各站的相关总的来说比与南区相关为好，说明平均联系法较为符合气候实际。

10区方案中，华德法划分各区所含站数比较均匀（表 1），但这不是气候区划的要求。它与平均联系法的分区有两处不同：华德法把长江下游分成两区，平均联系法把济南单划成一区。表面看来，把一个站划成一区似乎与其他区大小很不协调，但是，计算的相关系数证明这个分区较为合理(表 2,3)。

表 1 三种聚类方法分区所含站数

每区站数 / 方法 / 区号	5区方案			10区方案		
	平均联系法	华德法	重心法	平均联系法	华德法	重心法
1	8	7	46	7	7	35
2	9	10	1	1	6	1
3	14	14	1	6	4	3
4	14	14	1	3	8	4
5	5	5	1	8	6	1
6				6	6	2
7				11	5	1
8				3	3	1
9				2	2	1
10				3	3	1
站数均方差	3.9	4.1	20.1	3.1	1.9	10.6

表 2 济南与各站夏季降水距平百分率的相关系数

华北各站	大同	北京	唐山	沧州	石家庄	太原	榆林	
相关系数	0.14	0.22	0.27	0.56	0.36	0.32	0.26	
南区各站	临汾	安阳	莱阳	临沂	郑州	郎县	南阳	徐州
相关系数	0.29	0.45	0.30	0.22	0.25	−0.16	−0.12	0.30

表 3 长江下游南北两区各站间相关系数

南 \ 北	蚌埠	合肥	南京	上海	宁波	杭州
金华	0.39	0.68	0.64	0.72	0.78	0.86
屯溪	0.35	0.65	0.68	0.75	0.72	0.87
南昌	0.31	0.56	0.58	0.57	0.63	0.71
上饶	0.32	0.51	0.51	0.55	0.59	0.70
建阳	0.31	0.40	0.40	0.31	0.42	0.49

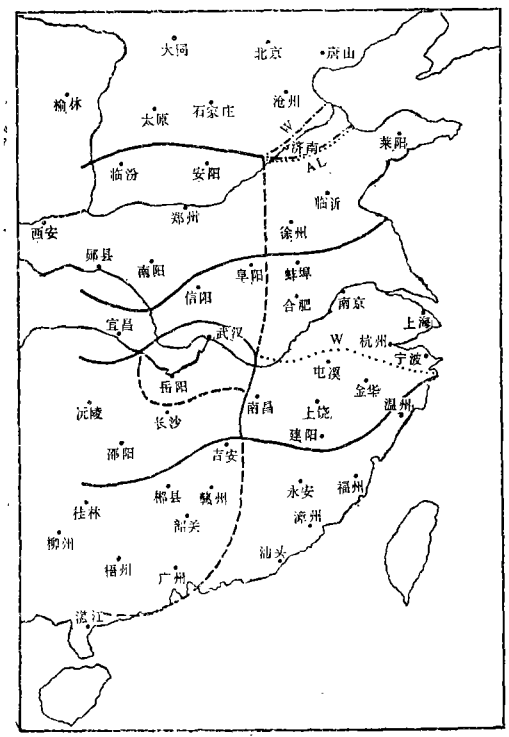

图 1 平均联系法和华德法对东部大陆夏季降水年际变化同步性的分区
(实线为五区方案两法共同的界线，点划相间的线为两法不同的界线，W 表示华德法，AL 表示平均联系法。实线加虚线为十区方案共同的界线，点线为两法不同的界线)

表 2 所示济南与其周围各站的相关系数，除沧州外均不超过 0.50；表 3 所示长江下游南北两区各站之间的相关系数，除位于最南端的建阳及最北端的蚌埠与各站的相关系数不到 0.50 以外，其他都在 0.50 以上。可见，平均联系法的划分较为准确。

四、结　　论

在气候分析中一直受到广泛应用的华德法，有优先合并小组的偏向，最后产生大小均匀的组，这并不反映气候现实。用于气候分区，小区域的特性得不到体现；用于划分气团，稀遇的气团不能识别。重心法有"滚雪球"效应，产生一个特大的组和很多极小的组，大的组包含了庞杂的成份，小的组往往是各别极端状态的罗列，起不到聚类所要求的作用。平均联系法则在很大程度上避免上列两法的偏向，产生较为符合气候现实的组。这些组的组内方差较小，即组内各成员的性质比较接近；组间的方差较大，各组有明显的差别。它的偏歧是各组的组内方差相差不大，这是否会影响分析结果，视具体对象而定。

参 考 文 献

[1] Johnston, R. J., Choice of classification: the subjectivity of objective methods, *Annals of the Association of American Geographers*, **58**, 575—589, 1968.

[2] Balling, R. C., Classification in climatology, *Spatial Statistics and Models*, edited by Gary Gaile and Cort Willmot, Redel Pub. Co. 1984.

[3] Anderberg, M. R., *Cluster analysis for applications*, Academic Press,New York & London, 1973.

[4] SAS Institute Lnc., *SAS User's Guide: Statistics*, 1982 edition, Cary, North Carolina, 1983.

[5] Willmott, C. J. and M. T. Vernon, Solar climates of the conterminous United States: A preliminary investigation, *Solar Energy*, **24**, 295—303, 1980.

[6] Willmott, G. J., P-mode principal components analysis, Grouping a nd precipitation regions in California, *Arch. Meteor. Geoph. Biokl.*, Ser. B, **24**, 257—267, 1978.

[7] Winkler, J. A., Regionalization of the diurnal distribution of summer precipitation, *Preprints of Sixth Conference on Hydrometeorology*, American meteorological society, Indianapolis, Indiana, U. S. A., Oct. 29—Nov. 1, 1985.

[8] Kalkstein, L. S., and P. Corrigan, A synoptic climatological approach for geographical analysis: Assessment of sulfur dioxide concentrations, *Ann. Assn. Am. Geog.*, **76**, 381—395, 1986.

[9] Ward, J. H. Jr., Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association*, **58**, 236—244, 1963.

[10] Lankford, P. M., and R. K. Semple, Classification and geography, *Geographia Polonica*, **25**, 7—30, 1973.

[11] Landford, P. M., Regionalization: Theory and alternative algorithms, *Geographical Analysis*, 1, 196—212, 1969.

[12] Zupan, J., *Clustering of Large Data Setr*, Research studies press, 1982.

[13] Kalkstein, L. S., G. Tan (谭冠日)and J. A. Skindlov, An evaluation of three clustering procedures for use in synoptic climatological classification, *Journal of climate and applied meteorology*, **26**, 717—730,1987.

[14] Williams, W. T., and G. N. Lance, Logic of computed-based intrinsic classifications, *Nature*, **207**, 159—161, 1965.

[15] Johnston, R. J., Regarding the delimitation of regions according to climatic fluctuations, *Archiv. Meteo. Geoph. und Biokl.*, Ser. B, **29**, 215—228, 1981.

[16] Boyce, A. J., Mapping diversity, A comparative study of some numerical methods, *Numerical Taxonomy*, edited by A. J. Cole, Academic press, London, 1—31, 1969.

[17] G. Tan (谭冠日), The patterns of summer dryness and wetness in eastern China, *Physical Geography*, **7**, 53—71, 1986.

EVALUATION OF CLUSTERING PROCEDURES IN CLIMATIC CLASSIFICATION

Tan Guanri

(Department of Atmospheric Sciences, Zhongshan University)

Abstract

Theoretical analyses on three clustering procedures show that: being influenced by the group size of merger at any step of fusion, Ward's method gives the priority to the merger with less observations, and finally produces clusters of close size. On the contrary, centroid method gives the priority to the merger with more observations, and finally produces a huge cluster and many small clusters some of which contain only one observation. Average linkage method is almost not influenced by the group size and produces cohesive clusters with close variance. It duplicates climatic reality with the greatest accuracy. A regionalization of year-to-year variation in summer precipitation in eastern China verifies above mentioned conclusions.