Statistics and Application 统计学与应用, 2021, 10(1), 92-105
Published Online February 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.101010

Hans汉斯

云南省各市经济发展的聚类分析

刘一帆

云南财经大学，云南 昆明
Email: liu_yifan6@163.com

收稿日期：2021年1月23日；录用日期：2021年2月16日；发布日期：2021年2月23日

摘　要

云南省地处我国西南边陲，其优越的地理位置、气候与资源条件为其经济发展带来了动力和机遇。本文通过采用类平均法、离差平方和法、k均值法对云南省的8个市和8个少数民族自治州的经济发展水平进行聚类分析。将云南省的16个地区按照经济综合实力分别划分为经济发展好、经济发展较好，经济发展一般，经济发展较差四个不同的类别。可为后续研究指导云南省各地区间的经济发展提供相应的理论依据。

关键词

经济发展，聚类分析，类平均法，离差平方和法，k均值法

Cluster Analysis of Economic Development in Yunnan Province

Yifang Liu

Yunnan University of Finance and Economics, Kunming Yunnan
Email: liu_yifan6@163.com

Received: Jan. 23rd, 2021; accepted: Feb. 16th, 2021; published: Feb. 23rd, 2021

Abstract

Yunnan province is located in the southwest border area of China. Its superior geographical location, climate and resource conditions have brought impetus and opportunity to its economic development. In this paper, the clustering analysis of the economic development level of 8 cities and 8 autonomous prefectures of ethnic minorities in Yunnan province is carried out by using the method of class average, sum of squares deviation and K-mean method. According to the comprehensive economic strength, the 16 regions in Yunnan province are divided into four different categories: good economic development, good economic development, general economic development

文章引用: 刘一帆. 云南省各市经济发展的聚类分析[J]. 统计学与应用, 2021, 10(1): 92-105.
DOI: 10.12677/sa.2021.101010

and poor economic development. It can provide the corresponding theoretical basis for the follow-up study to guide the economic development of various regions in Yunnan Province.

Keywords

Economic Development, Cluster Analysis, Method of Quasi-Average, Sum of Squares of Deviation, K-Means Method

1. 引言

素有“彩云之南”之称的云南省，近些年来，经济发展迅速，尤其在旅游经济的发展上一直保持着突飞猛进的势头。云南省整体的经济发展虽是一直处于良好稳定的状态，但是各个地区间的经济水平依旧有着不小的差距。比如，2016 年云南省生产总值为 14,720 亿元，昆明市的生产总值为 4300.08 亿元，但是怒江的生产总值只有 126.46 亿元。因此，云南省的经济发展是不平衡的：表现在有些地区经济发展较为迅速，而有些地区经济发展则较为滞后。因此，我们有必要对云南省各地区的经济差异作进一步的探讨和研究。

2. 云南省各地区经济发展概况分析

2.1. 地区生产总值概况

通过分析图 1：2016 年云南省各地区的生产总值雷达图，我们可以发现：云南省各地区间的经济发展存在很大的差异。昆明市 2016 年生产总值高达 4300.08 亿元。仅次于昆明的是曲靖的生产总值为 1768.41 亿元，两市生产总值相差 2531.67 亿元。2016 年云南省生产总值比较低的两个地区分别是迪庆和怒江：迪庆 2016 年生产总值为 176.88 亿元，而怒江 2016 年生产总值只有 126.46 亿元—与这一年生产总值最高的昆明市相差了 4173.62 亿元，相差将近 33 倍，差距非常悬殊。

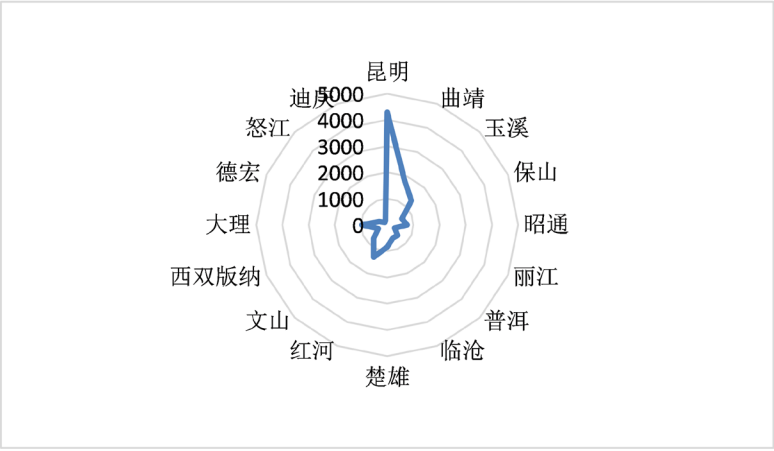

Figure 1. The total GDP profile of Yunnan province in 2016 (unit: 100 million yuan)
图 1. 2016 年云南省各地区生产总值概况(单位：亿元)

2.2. 地区居民消费概况

通过分析图 2：2016 年云南省各地区居民消费簇状柱形图，我们可以发现：2016 年云南省各个地区居民消费水平与各地区的生产总值整体上变化一致。昆明市居民消费水平依旧居于榜首，高达 1738.8 亿元，其次是曲靖市的居民消费为 706.82 亿元。由图可以看出，居民消费水平明显比较低的依旧是怒江和迪庆，迪庆 2016 年居民消费仅有 42.65 亿元，而怒江在这一年居民消费水平最低，只有 39.32 亿元，昆明市 2016 年的居民消费比怒江整整高出了 1699.48 亿元。

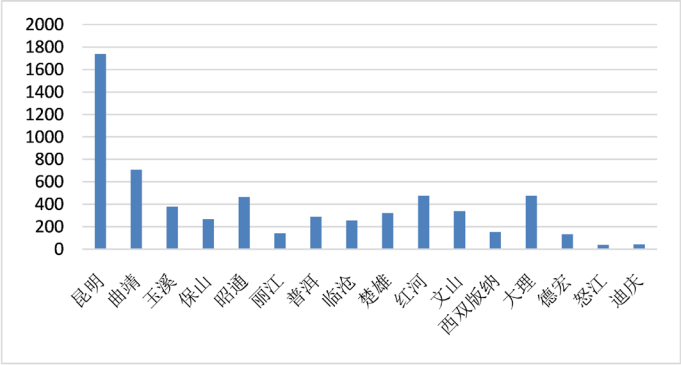

Figure 2. Overview of residents' Consumption in Yunnan Province in 2016 (unit: 100 million YUAN)
图 2. 2016 年云南省各地区居民消费概况(单位：亿元)

根据对云南省 2016 年各地区的生产总值和居民消费情况进行分析，我们大致可以看出，在 2016 年这一年，昆明市经济发展遥遥领先，曲靖市紧随其后。而怒江和迪庆的经济发展则一直处于较弱的水平。其余 12 个地区经济发展一般，在平均水平上下浮动。

3. 聚类分析

3.1. 指标选择与数据来源

在选择指标时，应充分考虑到所需数据的真实性及可获得性，同时应考虑各地区经济发展指标的相关性及全面性[1]。本文通过结合云南省的实际情况，综合考虑云南省 16 个地区的经济发展现状及影响因素，最终选定了 15 个反映云南省各地区经济发展综合情况的指标，作为本文选定的变量，为之后聚类分析提供可靠的数据支撑。具体变量如表 1 所示：

Table 1. Indicator system table
表 1. 指标体系表

变量	所表示的含义及单位	变量	所表示的含义及单位
X1	地区生产总值(亿元)	X9	职工平均工资(万元/人)
X2	人均 GDP (元/人)	X10	农村居民人均纯收入(元/人)
X3	第一产业产值(亿元)	X11	居民消费(亿元)
X4	第二产业产值(亿元)	X12	政府消费(亿元)
X5	第三产业产值(亿元)	X13	城乡居民储蓄存款额(万元/人)
X6	固定资产投资(亿元)	X14	进出口总额(亿元)
X7	公共财产收支总额(亿元)	X15	旅游总收入(亿元)
X8	社会消费品零售总额(亿元)		

本文使用的所有数据均来源于《2017 年云南省统计年鉴》[2]，原始数据见附录。

3.2. 采用类平均法进行聚类分析

类平均法有两种定义，一种定义方法是把类与类之间的距离定义为所有样品对之间的平均距离；另一种定义方法是类与类之间的平方距离为样品对之间平方距离的平均值；本文采用第二种定义方法，即

$$D_{KL}^2 = \frac{1}{n_K n_L} \sum_{i \in G_K, j \in G_L} d_{ij}^2 \tag{1}$$

类平均法可以较好地利用所有样品之间的信息[3]。

利用 SAS 程序进行类平方法聚类，因为样本数据单位不同，在聚类之前需要对各个变量数据作标准化变化。

SAS 程序如下：

```
PROC IMPORT OUT= WORK.test
DATAFILE= "C:\Users\DELL\Desktop\2016 年云南各地区经济指标.xls"
        DBMS=EXCEL REPLACE;
        RANGE="Sheet1$";
        GETNAMES=YES;
        MIXED=NO;
        SCANTEXT=YES;
        USEDATE=YES;
        SCANTIME=YES;
RUN;
proc cluster data=Work.Test method=ave std;
         id region;
proc tree horizontal;
         id region;
run;
```

聚类结果如下：

Average Linkage Cluster Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	11.0746184	9.3386761	0.7383	0.7383
2	1.7359423	1.0077605	0.1157	0.8540
3	0.7281818	0.1138740	0.0485	0.9026
4	0.6143078	0.1883527	0.0410	0.9435
5	0.4259551	0.2362695	0.0284	0.9719
6	0.1896856	0.0685247	0.0126	0.9846
7	0.1211609	0.0693247	0.0081	0.9927
8	0.0518362	0.0166281	0.0035	0.9961
9	0.0352081	0.0165828	0.0023	0.9985
10	0.0186253	0.0156333	0.0012	0.9997
11	0.0029921	0.0021748	0.0002	0.9999
12	0.0008173	0.0004237	0.0001	1.0000
13	0.0003936	0.0001181	0.0000	1.0000
14	0.0002755	0.0002755	0.0000	1.0000
15	-0.0000000		-0.0000	1.0000

The data have been standardized to mean 0 and variance 1

Root-Mean-Square Total-Sample Standard Deviation 1

Root-Mean-Square Distance Between Observations 5.477226

Figure 3. Correlation eigenvalue matrix of average method
图 3. 类平均法相关特征值矩阵

图 3 表示类平均法相关特征值矩阵，第一列代表协方差矩阵的特征值，第二列代表从上到下相邻两个特征值之差，第三列代表方差比，第四列代表方差累积比。

由于事先对样本数据进行了标准化处理，因此全部样本标准差的平方根为 1。观察值之间的均方根距离为 5.477226，代表变量之间也是较远距离的。

图 4 代表类平均法的聚类过程，从 NCL 可以看出 16 个变量一共聚类了 15 次。按照距离的远近，第一次聚类是普洱和文山聚为一类，因为两者之间标准化均方根距离最小，只有 0.11。以此类推，最后一次聚类是昆明和 CL2 聚成一个大类。

Cluster History

NCL	--Clusters Joined---		FREQ	Norm RMS Dist	Time
15	普洱	文山	2	0.11	
14	保山	临沧	2	0.1744	
13	CL15	楚雄	3	0.2353	
12	CL14	CL13	5	0.2622	
11	CL12	昭通	6	0.2867	
10	丽江	西双版纳	2	0.3204	
9	CL11	大理	7	0.3918	
8	曲靖	红河	2	0.4183	
7	CL9	CL10	9	0.4746	
6	CL7	德宏	10	0.519	
5	CL8	玉溪	3	0.6338	
4	CL6	怒江	11	0.6518	
3	CL5	CL4	14	0.7221	
2	CL3	迪庆	15	0.9364	
1	昆明	CL2	16	2.2762	

Figure 4. Clustering process of class average method
图 4. 类平均法聚类过程

类平均法最终得到图 5：类平均法树形图。若在坐标区间(0.70, 0.75)内切一刀，则将云南省 16 个地区分成四类。

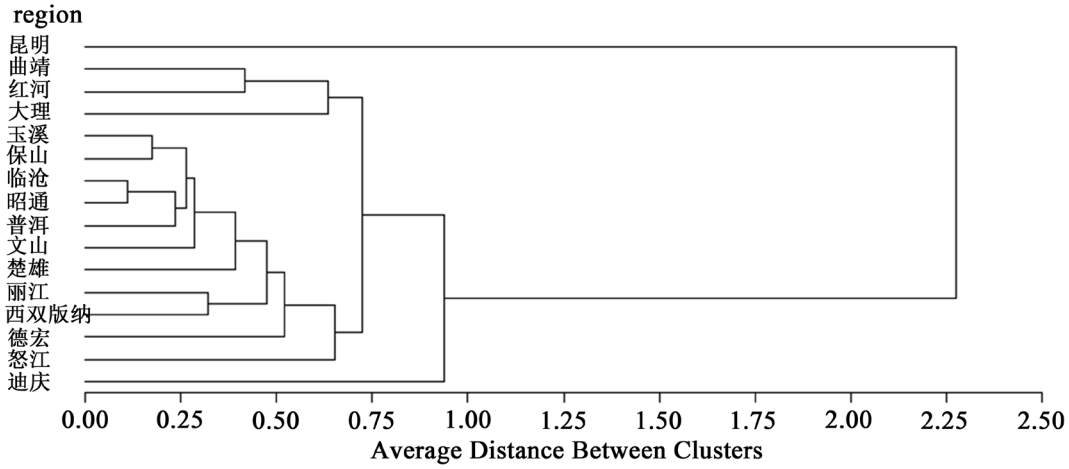

Figure 5. Class mean normal tree diagram
图 5. 类平均法树形图

第一类：昆明。

第二类：曲靖、红河、玉溪。

第三类：保山、临沧、普洱、文山、楚雄、昭通、大理、丽江、西双版纳、德宏、怒江。

第四类：迪庆。

3.3. 采用离差平方和法进行聚类分析

类中各样品到类重心的平方欧式距离之和称为(类内)离差平方和，设G_K和G_L合并成新类G_M，则G_K，G_L和G_M的离差平方和分别是：

$$W_K = \sum_{i \in G_K} \left(x_i - \bar{x}_K\right)' \left(x_i - \bar{x}_K\right) \tag{2}$$

$$W_L = \sum_{i \in G_L} \left(x_i - \bar{x}_L\right)' \left(x_i - \bar{x}_L\right) \tag{3}$$

$$W_M = \sum_{i \in G_M} \left(x_i - \bar{x}_M\right)' \left(x_i - \bar{x}_M\right) \tag{4}$$

对固定的类内样品数，它们反映了各自类内样品的分散程度；

G_K和G_L之间的平方距离为：

$$D_{KL}^2 = W_M - W_K - W_L \tag{5}$$

这种系统聚类法称为离差平方和法或 Ward 方法(Ward's method) [3]。

利用 SAS 程序进行离差平方和法聚类，因为样本数据单位不同，在聚类之前需要对各个变量数据作标准化变化。

SAS 程序如下：

```
PROC IMPORT OUT= WORK.test
DATAFILE= "C:\Users\DELL\Desktop\2016 年云南各地区经济指标.xls"
        DBMS=EXCEL REPLACE;
        RANGE="Sheet1$";
        GETNAMES=YES;
        MIXED=NO;
        SCANTEXT=YES;
        USEDATE=YES;
        SCANTIME=YES;
RUN;
proc cluster data=Work.Test method=war std nosquare;
          id region;
proc tree horizontal;
          id region;
run;
```

聚类结果如下：

图 6 代表离差平方和法的相关特征值矩阵，我们可以发现与离差平方和法的相关特征值矩阵结果一致，由此可以验证程序算法正确。

图 7 代表离差平方和法的聚类过程，从 NCL 可以看出 16 个变量一共聚类了 15 次。其聚类过程与类平均法聚类过程一致。

The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	11.0746184	9.3386761	0.7383	0.7383
2	1.7359423	1.0077605	0.1157	0.8540
3	0.7281818	0.1138740	0.0485	0.9026
4	0.6143078	0.1883527	0.0410	0.9435
5	0.4259551	0.2362695	0.0284	0.9719
6	0.1896856	0.0685247	0.0126	0.9846
7	0.1211609	0.0693247	0.0081	0.9927
8	0.0518362	0.0166281	0.0035	0.9961
9	0.0352081	0.0165828	0.0023	0.9985
10	0.0186253	0.0156333	0.0012	0.9997
11	0.0029921	0.0021748	0.0002	0.9999
12	0.0008173	0.0004237	0.0001	1.0000
13	0.0003936	0.0001181	0.0000	1.0000
14	0.0002755	0.0002755	0.0000	1.0000
15	-0.0000000		-0.0000	1.0000

The data have been standardized to mean 0 and variance 1

Root-Mean-Square Total-Sample Standard Deviation 1

Figure 6. Eigenvalue matrix associated with the sum of deviation squares method
图 6. 离差平方和法相关特征值矩阵

Cluster History

NCL	--Clusters Joined---		FREQ	SPRSQ	RSQ	False BSS	Tie
15	普洱	文山	2	0.0008	.999	0.3012	
14	保山	临沧	2	0.0020	.997	0.4775	
13	CL15	楚雄	3	0.0047	.993	0.7585	
12	丽江	西双版纳	2	0.0068	.986	0.8775	
11	昭通	CL13	4	0.0068	.979	0.8868	
10	CL14	CL11	6	0.0074	.971	0.979	
9	红河	大理	2	0.0110	.960	1.1124	
8	曲靖	CL9	3	0.0197	.941	1.4745	
7	CL12	德宏	3	0.0192	.922	1.5019	
6	CL8	玉溪	4	0.0316	.890	1.917	
5	怒江	迪庆	2	0.0335	.856	1.9403	
4	CL10	CL7	9	0.0369	.820	2.5066	
3	CL4	CL5	11	0.0804	.739	3.5344	
2	CL6	CL3	15	0.1150	.624	4.2585	
1	昆明	CL2	16	0.6241	.000	10.222	

Figure 7. Clustering process of the sum of deviation squares method
图 7. 离差平方和法聚类过程

在分成4类之前的并类过程中，RSQ的减少使逐渐进行的，改变不大。RSQ分成4类时，RSQ = 0.820，而下一次合并后分成3类时，RSQ下降较多，此时，RSQ = 0.739。通过分析RSQ统计量可知此时可以分成4类。同理，当RSQ分成3类时，RSQ = 0.739，而下一次合并后分成2类时，RSQ下降较多，此时，RSQ = 0.624。通过分析RSQ统计量可知此时也可以分成3类。

下面根据离差平方和法树形图，分别对这两种分类方式进行讨论。

最终得到图8离差平方和法树形图。

若在坐标区间(3.5, 4)内切一刀，则分成三类，它们分别是：

第一类：昆明。

第二类：曲靖、红河、大理、玉溪。

第三类：保山、临沧、昭通、普洱、文山、楚雄、丽江、西双版纳、德宏、怒江、迪庆。

离差平方和法这种聚类方法与类平均法的聚类结果大体上相一致。区别是，类平均法将玉溪与保山、临沧、昭通、普洱、文山、楚雄、丽江、西双版纳、德宏、怒江划分为了一类。将迪庆归为第四类。相比而言，类平均法的聚类方法要更加精准一些。

若在坐标区间(2.5, 3)内切一刀，则分为四类，它们分别是：

第一类：昆明。

第二类：曲靖、红河、大理、玉溪。

第三类：保山、临沧、昭通、普洱、文山、楚雄、丽江、西双版纳、德宏。

第四类：怒江、迪庆。

理想的聚类结果应该是类与类之间的特征明显不同而类内的特征彼此接近。根据我们之前对云南省各地区生产总值和居民消费的相关分析，与前两种聚类方式相比，这种聚类方式明显更贴近实际情况，更加吻合现实的经济意义。

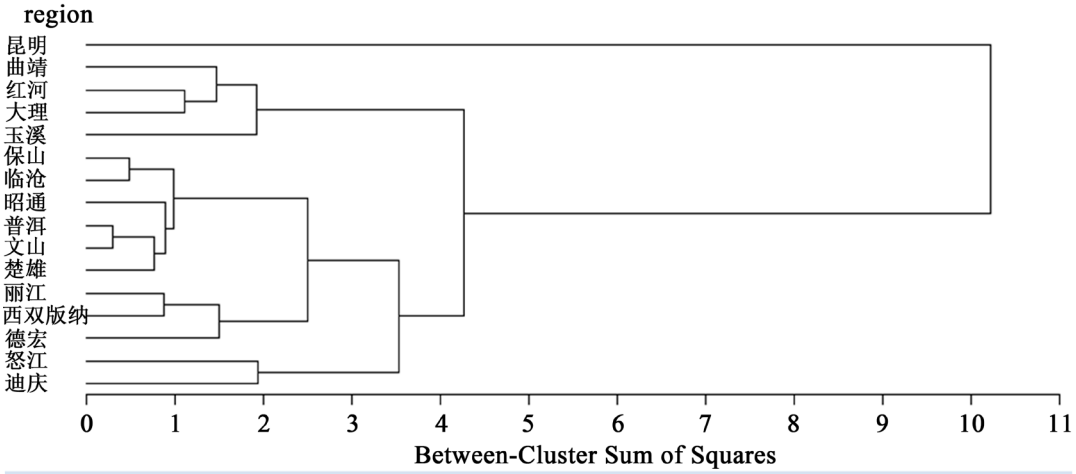

Figure 8. Normal tree of the sum of deviation squares
图 8. 离差平方和法树形图

3.4. 采用 K 均值法进行聚类分析

K 均值法的基本步骤为：

1) 选择 K 个样本作为初始凝聚点，或者将所有样品分成 k 个初始类，然后将这 k 个类的重心(均值)作为初始凝聚点。

2) 对所有的样品逐个归类，将每个样品归入凝聚点离它最近的那个类，该类的凝聚点更新为这一类目前的均值，直至所有样品归类。

3) 重复步骤(2)，直至所有的样品都不能再分配为止[3]。

接下来对这组数据采用 K 均值法聚类，根据之前的类平均法和离差平方和法的聚类结果，指定 K 均值法所允许的最大分类个数为 4 类，即“maxc = 4”。在聚类前同样对数据进行标准化变换。

SAS 程序如下：

```
PROC IMPORT OUT= WORK.test
DATAFILE= "C:\Users\DELL\Desktop\2016 年云南各地区经济指标.xls"
        DBMS=EXCEL REPLACE;
        RANGE="Sheet1$";
        GETNAMES=YES;
        MIXED=NO;
        SCANTEXT=YES;
        USEDATE=YES;
```

```
        SCANTIME=YES;
RUN;
proc standard data=Work.Test mean=0 std=1 out=stan;
proc fastclus data=stan maxc=5 drift list;
        var x1-x15;
        id region;
run;
```

聚类结果如下：

图 9 代表 k 均值法的初始凝聚点，最大聚类为 4 类，最大迭代为 1 次。将 15 个指标分为 4 个初始类，然后将这 4 个类的均值作为其初始凝聚点。输出结果如图 9 所示。

The FASTCLUS Procedure
Replace=FULL Drift Radius=0 Maxclusters=4 Maxiter=1

Initial Seeds

Cluster	x1	x2	x3	x4	x5	x6	x7	x8
1	3.347167385	2.562115562	0.726369203	3.202859888	3.558210999	3.115810503	3.255992342	3.624964337
2	0.823938126	-0.084866436	2.391573781	0.760116875	0.578063530	0.874317071	0.649966902	0.384863198
3	-0.762285617	0.980853695	-1.605533719	-0.754324286	-0.585553687	-0.684492234	-0.823883339	-0.578819583
4	-0.812537514	-0.528487943	-1.499369992	-0.819384758	-0.643667200	-0.889593727	-0.983271432	-0.603513211

Initial Seeds

Cluster	x9	x10	x11	x12	x13	x14	x15
1	0.356985836	1.895130156	3.361710866	3.334659245	3.579368996	2.990149010	3.015230262
2	-1.117335156	0.663031786	0.791937319	0.025956159	0.343433230	-0.341315443	-0.521675210
3	3.325155253	-1.201826987	-0.861938237	-0.944510540	-0.684915580	-0.682114732	-0.319717826
4	-0.023799056	-2.215263301	-0.870230399	-1.104746381	-0.706159391	-0.679361910	-0.970238849

Figure 9. Normal tree of the sum of deviation squares
图 9. 离差平方和法树形图

在图 10 聚类汇总信息图中，可以看到每一类的具体信息。例如，聚类 2 中共有 10 个观测值，RMS (Root Mean Square，简称均值平方根)的标准差为 0.4868，从该聚类种子到其余观测的最大距离为 2.8919。距离聚类 2 最近的类是聚类 4，二者之间的距离为聚类质心间的距离，即为 2.5832。

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	1	.	0		2	11.6855
2	10	0.4868	2.8919		4	2.5832
3	1	.	0		4	3.9640
4	4	0.5428	2.3540		2	2.5832

Figure 10. Cluster summary
图 10. 聚类汇总

图 11 变量的统计量信息图则给出了变量相关的统计量。观察可以发现，由于事先对变量进行了标准化处理，所以所有指标变量的 Total STD (总标准差)均为 1。对于变量 x1 来说，其 Within STD (类内合并标准差)为 0.35147，R-Square (相关系数)为 0.901177，拟合效果较好，RSQ/(1-RSQ) (类间方差同类内方差比)为 9.119058。

根据图表可以看出，大部分变量的拟合效果良好，但也有少部分变量拟合效果较差，如 x10、x2。

最终其 F 统计量值为 15.89。

The FASTCLUS Procedure
Replace=FULL Drift Radius=0 Maxclusters=4 Maxiter=1

Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
x1	1.00000	0.35147	0.901177	9.119058
x2	1.00000	0.73866	0.563500	1.290950
x3	1.00000	0.66783	0.643200	1.802693
x4	1.00000	0.42527	0.855314	5.911505
x5	1.00000	0.26339	0.944501	17.018392
x6	1.00000	0.48586	0.811150	4.295196
x7	1.00000	0.33620	0.909573	10.058645
x8	1.00000	0.19392	0.969917	32.241851
x9	1.00000	0.48298	0.813381	4.358514
x10	1.00000	0.87043	0.393885	0.649853
x11	1.00000	0.30376	0.926185	12.547360
x12	1.00000	0.26678	0.943063	16.563420
x13	1.00000	0.23281	0.956641	22.063470
x14	1.00000	0.64557	0.666593	1.999336
x15	1.00000	0.62701	0.685487	2.179517
OVER-ALL	1.00000	0.50137	0.798905	3.972762

Pseudo F Statistic = 15.89

Figure 11. Statistics for variables
图 11. 变量的统计量

图 12 给出了各变量在每一类的均值，例如，对于变量 x1 来说，其在聚类 1 中的均值为 3.347167385，在聚类 2 中的均值为 0.004784370，在聚类 3 中的均值为−0.762285617，在聚类 4 中的均值为−0.170254784。

Cluster Means

Cluster	x1	x2	x3	x4	x5	x6	x7	x8
1	3.347167385	2.562115562	0.726369203	3.202859888	3.558210999	3.115810503	3.255992342	3.624964337
2	0.004784370	-0.224101273	0.493326874	0.037179493	-0.088976459	0.038008193	0.061760061	-0.109273020
3	-0.762285617	0.980853695	-1.605533719	-0.754324286	-0.585553687	-0.684492234	-0.823883339	-0.578819583
4	-0.658181368	-0.325489131	-1.013526057	-0.705082634	-0.520723180	-0.702850051	-0.762427404	-0.488353639

Cluster Means

Cluster	x9	x10	x11	x12	x13	x14	x15
1	0.356985836	1.895130156	3.361710866	3.334659245	3.579368996	2.990149010	3.015230262
2	-0.300112195	0.105216952	0.021615339	0.051333119	-0.090839015	-0.242709380	-0.320094383
3	3.325155253	-1.201826987	-0.861938237	-0.944510540	-0.684915580	-0.682114732	-0.319717826
4	-0.170254784	-0.436368173	-0.678981506	-0.725869975	-0.496515818	0.029764881	0.126357849

Figure 12. Intra-class means of each variable
图 12. 各变量类内均值

图 13 给出了各变量在每一类中的标准差，例如，对于变量 x1 来说，其在聚类 2 中的标准差为 0.401227668，在聚类 4 中的标准差为 0.105670245。因为聚类 1 只有昆明市一个变量，聚类 3 中只有迪庆一个变量。所以在聚类 2、3 中并未输出其标准差。

Cluster Standard Deviations

Cluster	x1	x2	x3	x4	x5	x6	x7	x8
1	.	.	.	.	.	.	.	.
2	0.401227668	0.837860162	0.736147679	0.488493313	0.300078122	0.555774799	0.376076958	0.219309239
3	.	.	.	.	.	.	.	.
4	0.105670245	0.276536024	0.397816662	0.086915139	0.085753185	0.132646953	0.166833220	0.078250720

Cluster Standard Deviations

Cluster	x9	x10	x11	x12	x13	x14	x15
1	.	.	.	.	.	.	.
2	0.534880720	0.640760800	0.342735829	0.264225009	0.256320986	0.386843530	0.472178654
3	.	.	.	.	.	.	.
4	0.273500235	1.341212035	0.129120254	0.274295514	0.140327131	1.103671889	0.950635486

Figure 13. Intra-class standard deviations of each variable
图 13. 各变量的类内标准差

由图 14 聚类列表我们可以看到每一个地区所属的具体类别，以及其与种子的距离。例如，对于昆明市来说，在 k 均值法聚类中划分为第一类，其与种子的距离为 0。

Cluster Listing

Obs	region	Cluster	Distance from Seed
1	昆明	1	0
2	曲靖	2	2.8919
3	玉溪	2	2.8677
4	保山	2	1.3330
5	昭通	2	1.5669
6	丽江	4	1.3509
7	普洱	2	1.3124
8	临沧	2	1.3404
9	楚雄	2	0.9699
10	红河	2	1.7455
11	文山	2	1.1943
12	西双版纳	4	1.6654
13	大理	2	1.4955
14	德宏	4	1.7659
15	怒江	4	2.3540
16	迪庆	3	0

Criterion Based on Final Seeds = 0.4342

Figure 14. Cluster list
图 14. 聚类列表

最终由 k 均值法得到的聚类结果和之前的类平均法、离差平方和法的结果相比有很大区别。K 均值法聚类结果如下：

第一类：昆明。

第二类：曲靖、玉溪、保山、昭通、普洱、临沧、楚雄、红河、文山、大理。

第三类：迪庆。

第四类：丽江、西双版纳、德宏、怒江。

此程序中初始凝聚点的选择对于异常点很敏感，聚类后异常点很有可能单个地自成一类，例如昆明和迪庆，可能由于这两个地区城镇居民的消费结构与其他地区相比有一定的特殊性。

4. 结果分析

通过对比分析类平均法、离差平方和法、k 均值法的聚类结果，可以发现离差平方和法划分成 4 类的结果具有比较强的现实意义。

离差平方和法若在坐标区间(2.5, 3)内切一刀，划分为 4 类的结果是：

第一类：昆明。

无论以哪种聚类方式，昆明都可以单独的划分为“经济水平好”这一类，与之前的分析相一致：昆明经济发展相对于云南省其他地区来说遥遥领先。昆明市作为云南省省会，占据着优越的地理位置，加之其得天独厚的气候条件，使其具有云南省其他地区无法比拟的有利条件，这些方面都使昆明拥有一定的经济基础和发展优势。

第二类：曲靖、红河、大理、玉溪。

这四个地区经济发展都比较好。曲靖、红河以及玉溪都毗邻省会昆明，占据一定的地理优势，它们的经济结构和发展特点都比较类似。在 2016 年这一年，曲靖的生产总值高达 1768.41 亿元，红河的生产

总值为1333.79亿元，玉溪的生产总值为1311.88亿元，大理的生产总值为972.2亿元。这四个地区的生产总值在这一年仅居于昆明之后，排在第2~5的位置。排名比较集中，因此，再综合其他经济发展指标，将这四个地区划分为“经济水平较好”一类。

第三类：保山、临沧、昭通、普洱、文山、楚雄、丽江、西双版纳、德宏。

这九个地区经济发展相较于云南省其他地区来说，经济总量处于一般的水平。这些地区不具备较强的经济实力，贸易也处于中等水平，工业基础相对薄弱，人均收入也不算太高，但具有丰富的自然资源。因此在聚类时，将这九个地区归为“经济发展一般”这一类。

第四类：怒江、迪庆。在之前的分析中，2016年无论是其地区生产总值，还是居民消费水平，怒江和迪庆都处于较差的水平。怒江和迪庆都处于云南省的偏远地区，交通闭塞，经济贸易仅局限于当地的买卖活动，这两个地区同样没有可以带动其经济发展的特色产业。在这两个地区中，部分的村落生活条件依旧比较原始。因此经济发展较差，结合2016年的经济数据，将怒江和迪庆归为了“经济水平较差”这一类。

5. 结论

本文通过研究2016年云南省各地区的经济数据，重点探究了其地区生产总值概况及居民消费概况，大致了解了2016年云南省各地区的经济发展情况。再充分考虑到各地区经济发展指标的相关性及全面性，建立了反映云南省各地区经济发展综合情况的指标体系。再通过SAS软件，运用三种聚类方式：分别是类平均法、离差平方和法以及K均值法，进行聚类分析。最终结合对云南省各地区2016年经济发展情况的分析，以及对这三种方式的聚类结果进行比较，发现采用离差平方和法进行聚类在本数据的应用上更贴近于现实实际情况。

本文最终依照离差平方和法的聚类结果，结合2016年云南省各地区的经济发展情况，将16个地区按照经济综合实力分别划分为：

经济发展好：昆明。

经济发展较好：曲靖、红河、大理、玉溪。

经济发展一般：保山、临沧、昭通、普洱、文山、楚雄、丽江、西双版纳、德宏。

经济发展较差：怒江、迪庆。

这四个不同的类别。

对于云南省来说，要想实现一个省份的经济大发展，并不仅仅靠一个地区的一枝独秀，要努力协调好该省市各个地区经济的联合发展，努力缩小各个地区之间的经济发展差异。希望本次研究的聚类结果能够为后续研究指导云南省各地区之间的经济发展提供一定的理论支撑。

参考文献

[1] 曾五一, 肖红叶. 统计学导论[M]. 北京: 科学出版社, 2006.

[2] 贾俊平. 统计学[M]. 北京: 中国人民大学出版社, 2018.

[3] 王学民. 应用多元统计分析[M]. 上海: 上海财经大学出版社, 2017.

附录

Table 1. Economic indicators of various regions in Yunnan Province
表 1. 云南省各地区经济指标

region	x1	x2	x3	x4	x5	x6	x7
昆明	4300.08	64156	200.51	1660.11	2439.46	3920.07	1218.4
曲靖	1768.41	29155	335.56	674.91	757.94	1794	526.46
玉溪	1311.88	55389	135.02	685.34	491.52	893.69	364.41
保山	612.39	23654	151.32	213.03	248.04	667.33	270.16
昭通	765.53	14040	149.44	322.06	294.03	739.3	469.86
丽江	309.29	24116	47.34	120.34	141.61	345.63	198.32
普洱	567.54	21737	152.26	195.4	219.88	502.1	295.78
临沧	550.82	21906	154.67	185.77	210.38	917.04	251.66
楚雄	846.72	30948	162.63	321.95	362.14	1009.45	308.85
红河	1333.79	28588	214.3	601.33	518.16	2106.68	511.89
文山	735.88	20362	155.58	262.64	317.66	627.36	326.92
西双版纳	366.03	31338	92.22	98.61	175.2	415.07	147.79
大理	972.2	27360	205.32	370.94	395.94	750.61	376.83
德宏	323.55	25150	78.05	79.06	166.44	310.57	166.87
怒江	126.46	23289	20	37.87	68.59	120.92	92.81
迪庆	176.88	43247	11.39	64.11	101.38	315.46	135.13
region	x8	x9	x10	x11	x12	x13	x14
昆明	2310.09	6.84	12555	1738.8	433.89	4124.21	66.81
曲靖	564.97	5.33	10380	706.82	126.84	1070.12	6.3
玉溪	326.77	6.04	11968	379.04	141.34	752.34	20.19
保山	200.18	5.32	9426	267.86	127.59	518.14	2.66
昭通	237.63	5.97	7951	464.94	129.64	647.39	0.06
丽江	104.85	6.43	8750	141.91	64.63	330.98	0.69
普洱	163.04	6.6	8669	289.38	145.11	442.78	11.21
临沧	173.65	5.92	8914	256.65	83.89	305.6	6.65
楚雄	298.77	6.84	9181	319.92	88.72	563.39	5.11
红河	366.56	6.31	9449	476.31	143.23	989.68	20.16
文山	323.72	6.66	8403	339.12	153.25	517.12	5.93
西双版纳	116.41	6.44	11049	151.66	59.26	339.65	8.79
大理	332.99	6.68	9612	474.67	152.34	795.96	2.64
德宏	124.73	5.88	8659	131.6	82.48	359.35	42.52
怒江	32.63	6.45	5299	39.32	21.91	79.51	0.16
迪庆	45.93	9.88	7088	42.65	36.78	99.56	0.11

Continued

region	x14	x15
昆明	66.81	1073.53
曲靖	6.3	153.04
玉溪	20.19	162.86
保山	2.66	173
昭通	0.06	147.06
丽江	0.69	608.76
普洱	11.21	168.36
临沧	6.65	112.07
楚雄	5.11	178.18
红河	20.16	274.61
文山	5.93	151.26
西双版纳	8.79	420.28
大理	2.64	534.58
德宏	42.52	221.43
怒江	0.16	36.3
迪庆	0.11	205.6