2007 年江苏省十三大市经济建设主要指标聚类分析

黄秋桥，姜楠楠，程猛
辽宁工程技术大学理学院，辽宁阜新 (123000)
E-mail：huangqiuqiao@126.com

摘　要：江苏省作为全国经济大省，对全国经济发展有着突出的贡献，但是江苏区域的经济发展的梯度特征却十分明显，苏北的经济发展水平远远落后于苏南地区。本文采用了多元统计分析方法中的谱系聚类分析，根据江苏十三个大市的各项经济指标划分区域，这对找出各市差距，促进江苏经济综合发展提供了理论依据。

关键词：谱系聚类法；江苏 GDP 经济建设；经济发展状况
中图分类号：O212.4

1　引 言

近年来，江苏省的经济发展一直很快速，但是与全国经济排名第一的广东省比较而言差距逐渐有拉大的趋势。究其原因不难发现江苏苏北与苏南地区发展严重不平衡。因此有必要研究一下江苏各市的经济发展状况，分析一下江苏区域经济发展的梯度特征，对各市的经济建设做出一个较为合理的分析，评价，这对江苏省经济全局统筹兼顾的发展有一定的指导作用。

本文以江苏省的 13 个大市研究对象，运用多元统计分析中的谱系聚类分析[2]，聚类分析是统计学中研究“物以类聚”问题的多元统计分析方法，在统计分析的应用领域已经得到了极为广泛的应用，是初始的、探索性的科学研究工作的基本方法之一。聚类对象间的间亲疏程度一般用相似的概念描述，而对象间的相似性是通过距离或相似系数度量的。聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。聚类分析的目标就是在相似的基础上收集数据来分类。聚类源于很多领域，包括数学，计算机科学，统计学，生物学和经济学。在不同的应用领域，很多聚类技术都得到了发展，这些技术方法被用作描述数据，衡量不同数据源间的相似性，以及把数据源分类到不同的簇中。作为从数值分类中分离出来的一个数学分支,它从数据分析的角度,给出在同一分类过程中始终如一的定量方法,从而避免了普通分类中主观随意性大的弊端,是一种更为准确、更为细致的科学分类工具。谱系聚类分析是聚类分析的一种。

2　指标体系的建立

本文在遵循科学性、合理性、可比性和可操作性的原则下，以江苏 13 个市为样本，选取了下列反映江苏区域经济发展的 6 项统计指标，建立江苏各市经济发展指标体系。数据来源—江苏省统计年鉴【2008】[1]如表 2-1。

X1——地区生产总值（亿元）

X2——人均生产总值增长率（%）

X3——固定资产投资（亿元）

X4——社会消费品零售总额（亿元）

X5——全部工业总产值（亿元）

X6——在岗职工平均工资(元）

表2–1 2007年江苏13大市影响经济状况的因素具体数据

指标 / 市区	地区生产总值（亿元）	人均生产总值增长率（%）	固定资产投资（亿元）	社会消费品零售总额（亿元）	全部工业总产值(亿元)	在岗职工平均工资（元）
南京市区	3015.72	115.4	1364.35	1296.41	5387.51	36672
无锡市区	2162.92	114.6	919.65	728.43	4422.22	35583
徐州市区	897.60	115.8	358.87	296.53	1093.96	30652
常州市区	1419.91	115.6	620.98	455.46	3398.59	32672
苏州市区	2295.29	115.4	786.37	580.59	4744.29	33295
南通市区	551.94	115.9	248.49	170.91	958.58	30154
连云港市区	274.64	112.5	197.86	101.51	411.35	25701
淮安市区	476.56	114.9	245.95	171.30	652.39	21627
盐城市区	387.89	112.7	217.81	142.99	630.70	23123
扬州市区	603.32	115.8	242.03	193.67	1080.36	29266
镇江市区	534.96	115.3	260.30	159.09	945.72	30438
泰州市区	325.39	115.7	176.33	90.55	703.60	24148
宿迁市区	205.16	117.3	119.96	60.14	179.07	21657

数据来源：《江苏省统计年鉴 2008》 网址 http://www.jssb.gov.cn/jstj/jsnj/2008/nj17.htm

3 谱系聚类分析模型的构建

3.1 谱系聚类分析模型

3.1.1 基本原理

谱系聚类法[3]是本文将采用的方法。其思路为：首先将每个数据对象各视为一类，根据类与类之间的距离或相似程度将最相似的类加以合并，再计算新类与其它类之间的相似程度，并选择最相似的类加以合并，这样每合并一次就减少一类，不断续这一过程，直到所有数据对象合并为一类为止。最后根据各类之间的亲疏关系，逐步画成一张完整的分类系统图，又称谱系图。其相似程度由距离或者相似系数定义。进行类别合并的准则是使得类间差异最大，而类内差异最小。

3.1.2 计算步骤

（1）数据标准化

本文采用极差正规化变换方法，变化如下：

$$x_{ij}^{*}=\frac{x_{ij}-\min\limits_{1\le t\le n}x_{ij}}{R_j}\qquad (i=1,2,\cdots,n;\ j=1,2,\cdots,m) \tag{1}$$

（2）计算 n 个样本的两两间距离[4]，得到样品间的距离矩阵 D。本文距离公式采用马氏距离，公式如下，其中 S 为样品协方差矩阵：

$$d_{ij}=\left[\left(x_i-x_j\right)^T S^{-1}\left(x_i-x_j\right)\right]^{\frac{1}{2}} \qquad (i=1,2,\cdots,n) \qquad (2)$$

(3)初始（第一步：$i=1$）n 个样品各自构成一类，类的个数 $k=n$，第 t 类 $G_t=\{X_{(t)}\}$ $(t=1,\cdots,n)$。此时类间的距离就是样品间的距离。然后对步骤 $i=2,\cdots,n$ 执行并类过程的步骤（3）和（4）。

(4)对步骤 i 得到的距离矩阵 $D^{(i-1)}$，合并类间距离最小的两类为新一类。此时类的总个数 k 减少 1 类，即

$$k=n-i+1 \qquad (3)$$

(5)计算新类与其他类的距离，得到的距离矩阵 $D^{(i)}$。若合并后类的总个数 k 仍大于 1，重复步骤（3）和（4）；直到类的总个数为 1 时转到步骤（5）。

(6)画谱系聚类图。

(7)决定分类的个数及各类的成员。

4 江苏十三市经济指标的实证分析

4.1 聚类分析模型具体分析

本文处理的是各个城市的各项经济建设指标，由于表 2-1 中的各变量之间存在不同量纲、不同数量级的情况，故在考虑类与类的距离时，会出现片面强调某些指标的重要性的情况，因此存在数据标准化的必要性，数据标准化的目的是使这些变量具有可比性，消除量纲的影响，使数据得以在更平等的条件下进行聚类和分析。

又由于各项指标间有部分相关性，因此采用马氏距离公式来求变量间的距离，达到消除相关性影响的目的，最后聚类时采用 linkage 的 average 未加权平均法完成。

4.2 MATLAB 具体实现

4.2.1 输入数据并标准化

代码如下：

```
X = [3015.72 115.4 1364.35 1296.41 5387.51 36672;2162.92 114.6 919.65 728.43 4422.22 35583;897.60 115.8 358.87 296.53 1093.96 30652;1419.91 115.6 620.98 455.46 3398.59 32672;2295.29 115.4 786.37 580.59 4744.29 33295;551.94 115.9 248.49 170.91 958.58 30154;274.64 112.5 197.86 101.51 411.35 25701;476.56 114.9 245.95 171.30 652.39 21627;387.89 112.7 217.81 142.99 630.70 23123;603.32 115.8 242.03 193.67 1080.36 29266;534.96 115.3 260.30 159.09 945.72 30438;325.39 115.7 176.33 90.55 703.60 24148;205.16 117.3 119.96 60.14 179.07 21657]
[MaxV, I]=max(X);  %求 X 矩阵每列的最大值
[MinV, I]=min(X);  %求 X 矩阵每列的最小值
```

```
[R,C]= size(X);  %求矩阵 X 的行数，列数
s=(X-ones(R,1)*MinV).*(ones(R,1)*(ones(1,C)./(MaxV-MinV)))  %极差正规化变换
```
公式

最后可得到标准化矩阵

```
s =
    1.0000    0.6042    1.0000    1.0000    1.0000    1.0000
    0.6966    0.4375    0.6426    0.5406    0.8147    0.9276
    0.2464    0.6875    0.1920    0.1912    0.1757    0.5999
    0.4322    0.6458    0.4026    0.3198    0.6181    0.7341
    0.7437    0.6042    0.5355    0.4210    0.8765    0.7755
    0.1234    0.7083    0.1033    0.0896    0.1497    0.5668
    0.0247         0    0.0626    0.0335    0.0446    0.2708
    0.0966    0.5000    0.1012    0.0899    0.0909         0
    0.0650    0.0417    0.0786    0.0670    0.0867    0.0994
    0.1417    0.6875    0.0981    0.1080    0.1730    0.5077
    0.1173    0.5833    0.1128    0.0800    0.1472    0.5856
    0.0428    0.6667    0.0453    0.0246    0.1007    0.1676
       0    1.0000         0         0         0    0.0020
```

4.2.2 计算距离矩阵

代码如下：

```
c = cov(s); %求 s 的协方差矩阵
a =s';  %求矩阵 s 的转置矩阵
for i = 1:13
  for j = 1:13
    d(i,j) = sqrt((a(:,i)-a(:,j))'*inv(c)*(a(:,i)-a(:,j)));  %马哈拉距离公
```
式

```
  end
end
D=squareform(d)  %拉直矩阵 d
```

得到距离矩阵

```
D =
  Columns 1 through 13
    3.7889    4.0892    4.2334    4.8896    3.8785    4.3579    3.8060
4.0885    4.1091    4.1977    3.8927    4.3450    3.4687
  Columns 14 through 26
    2.7921    3.6296    3.0096    2.8894    3.2387    3.8230    4.5274
2.3502    3.0809    3.8321    4.6484    3.9154    2.2937
  Columns 27 through 39
    3.4046    3.2599    3.9356    3.5692    2.2796    3.4651    3.5532
4.1441    2.7406    3.6551    3.7568    3.8029    3.4229
  Columns 40 through 52
```

```
       2.9406    2.2989    3.7182    3.9033    4.4148    3.7261    3.9537
4.0244    4.1876    3.7557    4.3463    2.8682    3.0922
      Columns 53 through 65
       3.2769    2.1424    1.2338    2.0556    3.0440    2.7584    1.8996
3.8564    2.4376    2.8649    4.1332    2.4557    3.8674
      Columns 66 through 78
       3.1102    1.8128    1.9862    3.1897    3.5156    2.6540    4.0366
3.3034    2.8133    4.0129    2.4278    3.2460    1.8698
```

4.2.3 创建聚类信息矩阵

代码如下：

```
G = linkage(D,'average')    %采用未加权平均距离法聚类
```

得到矩阵 G

```
G =
   6.0000   11.0000    1.2338
   8.0000   12.0000    1.8128
   7.0000    9.0000    1.8996
  13.0000   15.0000    1.9280
   3.0000   14.0000    2.2867
   2.0000    4.0000    2.7921
  10.0000   18.0000    3.0050
  16.0000   17.0000    3.1505
  20.0000   21.0000    3.2216
  19.0000   22.0000    3.4002
   5.0000   23.0000    4.0001
   1.0000   24.0000    4.1397
```

对矩阵 G 的解释如表 4-1：

表4–1 矩阵 G 的解释说明表

	参与合并的聚类 I	参与合并的聚类 II	类间距离
第一轮合并操作	6	11	1.2338
第二轮合并操作	8	12	1.8128
第三轮合并操作	7	9	1.8996
第四轮合并操作	13	15=[8，12]	1.9280
第五轮合并操作	3	14=[6,11]	2.2867
第六轮合并操作	2	4	2.7921
第七轮合并操作	10	18=[3,14]	3.0050
第八轮合并操作	16=[7,9]	17=[3,15]	3.1505
第九轮合并操作	20=[10,18]	21=[16,17]	3.2216
第十轮合并操作	19=[20,21]	22=[2,4]	3.4002
第十一轮合并操作	5	23=[19,22]	4.0001
第十二轮合并操作	1	24=[5,23]	4.1397

4.2.4 创建聚类信息矩阵

运用 MATLAB 软件中的 dendrogram(G)函数绘制谱系聚类图像：见图 4-2

代码如下：

```
dendrogram(G) %绘制谱系聚类图
```

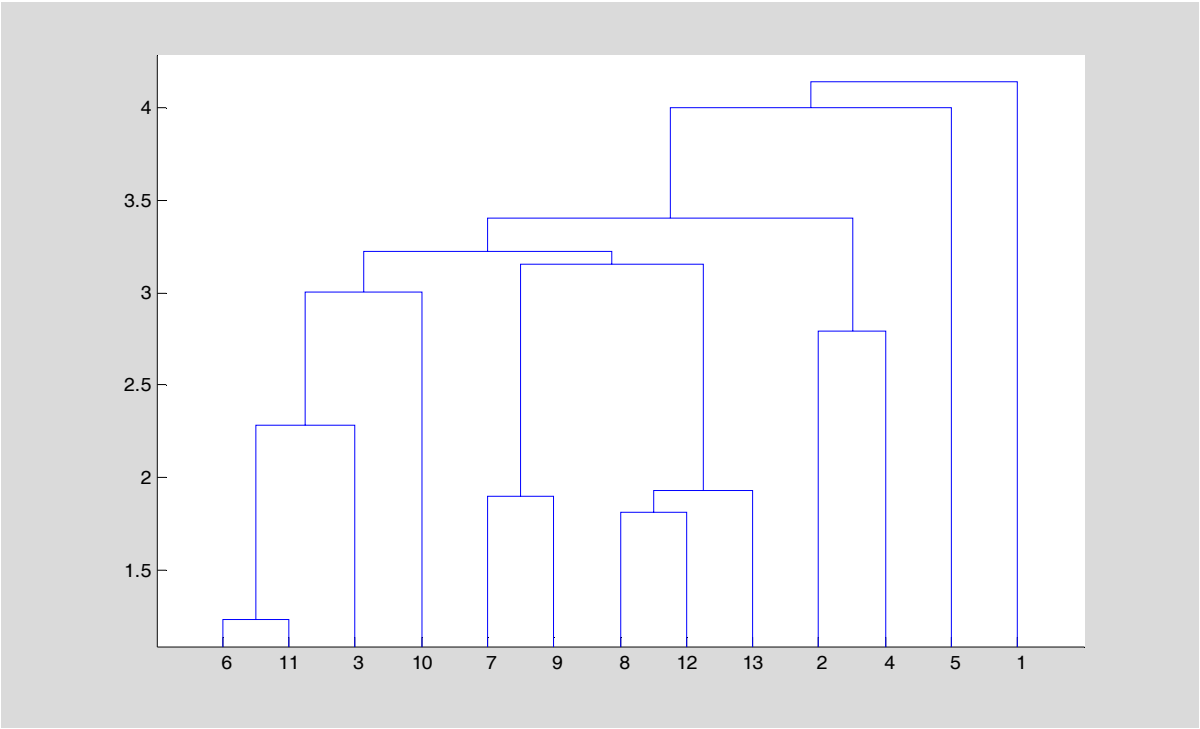

图4–2 2007年江苏13大市经济指标谱系聚类图

4.2.4 创建指定个数的类

引用 MATLAB 中的 cluster 指令的功能是根据 linkage 函数的输出 ans 创建指定个数的类，引用参数'maxclust' 控制输出信息使之给出分类的具体结果，输出 T 是一个列向量，自上而下按对象的编号从小到大顺次排列，T 中数字相同的对象归聚为一类，结果如下：

```
T = cluster(G,'maxclust',5) %采用 maxclust 方法将 G 分为 5 类
```

运用 MATLAB 软件计算，根据各大市的综合得分值进行分类，分类结果如下表 4-3：

表4–3 江苏13大市分类表

第一类	南京
第二类	苏州
第三类	无锡、常州
第四类	徐州、南通、扬州、镇江
第五类	连云港、淮安、盐城、泰州、宿迁

5 总结

由聚类分析的结果不难看出：

（1）江苏省 13 大市可分为 5 个梯队，第一梯队为南京，第二梯队为苏州，第三梯队为无锡、常州，第四梯队为徐州、南通、扬州、镇江，第五梯队为连云港、淮安、盐城、泰州、宿迁。

（2）江苏省的经济总体发展是快速的，在全国也是名列前茅的，但是地区发展的高度不平衡，严重阻碍了它的进一步高速发展。

近年来，江苏省会南京 GDP 发展首次超越苏州，体现了一省省会老大的优势和地位，苏州由于拉动外资企业和扩大内需趋近于饱和，步伐的放慢而名列第二。常州则加快了经济

发展的步调，与排名第三的无锡差距正在缓慢缩小。至于苏北三市近年发展也相当快，正在积极进取，努力改善自己的不利地位。但在国内生产总值、第三产业占 GDP 比重、全部工业总产值、社会消费品零售总额、地方预算内财政收入上，这三个市均处于落后状态。人们期待着苏北三市的腾飞

江苏省的实力是显而易见的。在 2003 年的全国各省市 GDP 排名当中，江苏省位列第二，在 2002 和 2003 年的全国城市 GDO 排名当中，前三梯队的苏州，无锡，南京都有位置。但是，我们也要看到一些不足之处。江苏省的 GDP 虽然发展迅速并且对全国 GDP 的贡献越来越大，但与第一名的广东省仍有一定的差距。而江苏省弱势群体——苏北三市与其它城市的差距又太大，可以说是扯了江苏省总体发展的后腿。

(3)所以从各方面看来，我省要继续快速发展，当务之急是要发展落后的城市，使其跟上先进的城市，而经济发达的城市更要努力发展，不能落后。在这里，比较好的而且已经在实施的方法就是发达的城市带动落后的城市，采用一对一的方式，互相扶持，取长补短，共同发展。如现在排名第二的苏州市就与末尾的宿迁市结成了对子，帮助该市发展，并已经取得了一定的成效。

参考文献

[1] 《江苏省统计年鉴 2008》
[2] 高惠旋.《应用多元统计分析》.北京.北京大学出版社,2005
[3] 朱永生.《实验数据多元统计分析》.北京:科学出版社, 2009
[4] 理查德.A.约翰逊,迪安.W.威克恩.《实用多元统计分析》.第 6 版.清华大学出版社,2008.11

In 2007 Jiangsu Province 13 leading market economic development major targets or quotas cluster analysis

Quiqiao Huang, Nannan Jiang, Meng Cheng
Liaoning engineering technology university college of science, Fuxin, Liaoning (123000)

Abstract

Jiangsu Province as national economy big province, has the prominent contribution to the national economic development, but the Jiangsu region's economic development's gradient characteristic is very actually obvious, the northern Jiangsu economic development level falls behind by far the Southern Jiangsu area. This article has used in the multi-dimensional statistical analysis method evolution cluster analysis, divides the region according to Jiangsu 13 leading market's each economic indicator, this to discovers various cities disparity, promoted the Jiangsu economy integrated development to provide the theory basis.

Keywords: Evolution cluster law; Jiangsu GDP economic development; economic development condition