第四章　时间序列分析与系统聚类分析

- 时间序列分析
- 系统聚类分析

§4.1 时间序列分析

时间序列，也叫时间数列或动态数列，时要素（变量）的数据按照时间序列而形成的一种数列，它反映了要素随时间变化的过程。

地理过程的时间序列分析，就是通过分析地理要素随时间变化的历史过程，揭示其发展变化规律，并对其未来状态进行预测。

时间序列分析的基本原理

（一）时间序列的组合成分

为分析时间序列的趋势模式，必须首先了解时间序列的组合成分（component）。**主要分：**

1. **长期趋势**（T）__是时间序列随时间的变化而逐渐增加或减少的长期变化之趋势。

2. **季节变动**（S）__是时间序列在一年中或固定时期内，呈现出的固定规则的变动。

3. **循环变动**（C）__是指沿着趋势线如钟摆般的变动。

4. **不规则变动**（I）---是指在时间序列中由于随机因素影响所引起的变动。

（二）时间序列的组合模型

统计学家根据时间序列四种成分的不同结合方法，而提出了时间序列的两种组合模型，即加法模型和乘法模型。

1. **加法模型**---加法模型假定时间序列是基于四种成分的相加而成的，其基本假设是：各成分彼此间独立，无交互影响，亦即长期趋势并不影响季节变动。

若 Y 表示时间序列，则加法模型为：

$$Y = T + S + C + I \qquad (4.1.1)$$

2. 乘法模型---乘法模型假定时间序列是基于四种成分的相乘而成的。在乘法模型中，各成分之间明显的存在相互依赖关系。该模型的基本方程为：

$$Y = T \times S \times C \times I \qquad (4.1.2)$$

趋势拟合方法

1. 平滑法

时间序列分析的平滑法主要有三类

◆**移动平均法**：设某一时间序列为 y_1，y_2，…，y_t，则下一期（t+1时刻）的预测值为

$$\hat{y}_{t+1}=\frac{1}{n}\sum_{j=0}^{n-1}y_{t-j}=\frac{y_t+y_{t-1}+\cdots+y_{t-n+1}}{n}=\hat{y}_t+\frac{1}{n}(y_t-y_{t-n}) \quad (4.1.3)$$

式中，$\hat{y}_t$ 为t点的移动平均值，n称为移动时距（点数）。

◆ 滑动平均法 ：其计算公式为：

$$\hat{y}_t = \frac{1}{2l+1}(y_{t-l} + y_{t-(l-1)} + \cdots + y_{t-1} + y_t + y_{t+1} + \cdots + y_{t+l}) \quad (4.1.4)$$

（4.1.2）式中， $\hat{y}_t$为t点的滑动平均值，L为单侧平滑时距（点数）。

若L=1，则以下公式式称为三点滑动平均，其计算公式为：

$$\hat{y}_t = (y_{t-1} + y_t + y_{t+1})/3 \quad (4.1.5)$$

若L=2，则（4.1.4）式称为五点滑动平均，其计算公式为：

$$\hat{y}_t = (y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2})/5 \qquad (4.1.6)$$

◆指数平滑法

① 一次指数平滑

$$\hat{y}_{t+1} = \sum_{j=0}^{n-1} \alpha(1-\alpha)^j y_{t-j} = \alpha y_t + (1-\alpha)\hat{y}_t \qquad (4.1.7)$$

α 为平滑系数。一般时间序列较平稳， α 取值可小一些[一般取 α ∈（0.05,0.3）]；若时间序列数据起伏波动比较大，则 α 应取较大的值[一般取 α ∈（0.7,0.95）]。

② 高次指数平滑法

一次指数平滑法不能跨期预测，对其进行改进，可以得到能够跨期预测的高次指数平滑法。令 $S_t^{(1)}$ 为一次指数平滑值，即

$$S_t^{(1)} = \alpha y_t + (1-\alpha) S_{t-1}^{(1)} \qquad (4.1.8)$$

对上式再作指数平滑，可得二次指数平滑值，即

$$S_t^{(2)} = \alpha S_1^{(1)} + (1-\alpha) S_{t-1}^{(2)} \qquad (4.1.9)$$

▲二次指数平滑法的预测公式为

$$\hat{y}_{t+k} = a_t + b_t T \qquad (4.1.10)$$

在（4.1.10）式中，T代表从基期t到预测期的期数，

$$a_t = 2S_t^{(1)} - S_t^{(2)} \qquad (4.1.11)$$

$$b_t = \frac{\alpha}{1-\alpha}\left(S_t^{(1)} - S_t^{(2)}\right) \qquad (4.1.12)$$

对二次指数平滑再作指数平滑，可得三次指数平滑公式：

$$S_t^{(3)} = \alpha S_1^{(2)} + (1-\alpha)S_{t-1}^{(3)}$$

▲三次指数平滑法的预测公式为

$$\hat{y}_{t+k} = a_t + b_t T + c_t T^2 \qquad (4.1.14)$$

（4.1.9）式中，$a_t = 3S_t^{(1)} - 3S_t^{(2)} + S_t^{(3)}$ （4.1.15）

$$b_t = \frac{\alpha}{2(1-\alpha)^2}\left[(6-5\alpha)S_t^{(1)} - 2(5-4\alpha)S_t^{(2)} + (4-3\alpha)S_t^{(3)}\right] \qquad (4.1.16)$$

$$c_t = \frac{\alpha^2}{2(1-\alpha)^2}\left[S_t^{(1)} - 2S_t^{(2)} + S_t^{(3)}\right] \qquad (4.1.17)$$

2. 趋势线法

分析时间序列的长期趋势，往往需要拟合一条适当的趋势线，用以概括的反映长期趋势的变化态势。最常用的趋势线有如下：

(1) 直线型趋势线，即 $y_t = a + bt$ ；

(2) 指数型趋势线，即 $y_t = ab^t$ ；

(3) 抛物线型趋势线，即 $y_t = a + bt + ct^2$ ；

§4.2 系统聚类分析

聚类分析，亦称群分析或点群分析，它是研究多个要素事物分类问题的数量方法。其基本原理是，根据样本自身的属性，用属性 方法按照某种相似性或差异性指标，定量的确定样本之间的亲疏关系，并按照这种关系的程度对样本进行聚类。

一，聚类要素的预处理

在研究中，因为不同的要素的数据往往具有不同的单位和量纲,进行聚类分析之前，首先要对聚类要素进行数据处理。假设有m 个聚类的对象，每一个聚类对象都有n个要素构成它们所对应的要素数据可用4.2.1给出。

聚类对象	要素					
	X_1	X_2	…	X_j	…	X_n
1	X_{11}	X_{12}	…	X_{1j}	…	X_{1n}
2	X_{12}	X_{22}	…	X_{2j}	…	X_{2n}
:	:	:	…	:	…	:
i	X_{i1}	X_{i2}	…	X_{ij}	…	X_{in}
:	:	:	…	:	…	:
m	X_{m1}	X_{mi}	…	X_{mj}	…	X_{mn}

（1）总和标准化。分别求出个聚类要素所对应的数据的总和，以各要素的数据处以该要素的数据的总和，即：

$$x_{ij}' = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \quad (i=1,2,...,m;\ j=1,2,...,n) \qquad (4.2.1)$$

这种标准化方法所得到的新数据 x_{ij}' 满足

$$\sum_{i=1}^{m} x_{ij}' = 1 \qquad (j=1,2,...,n)$$

（2）标准差标准化，即

$$x_{ij}' = \frac{x_{ij} - \overline{x}_j}{s_j} \quad (i = 1,2,...,m;\ j = 1,2,...,n) \qquad (4.2.2)$$

由这种标准化方法所得的新数据 x_{ij}'，各要素的平均值为0，标准差为1。

(3) 极大值标准化，即

$$x_{ij}' = \frac{x_{ij}}{\max_i \{x_{ij}\}} \quad (i = 1,2,...,m;\ j = 1,2,...,n) \qquad (4.2.3)$$

极差的标准化，即

$$x'_{ij} = \frac{x_{ij} - \min\limits_{i}\{x_{ij}\}}{\max\limits_{i}\{x_{ij}\} - \min\limits_{i}\{x_{ij}\}} \quad (i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n) \qquad (4.2.4)$$

进过这种变化所得的新数据，各要素的极大值为1，极小值为0，其余的数值均在0与1之间。

二，距离的计算

距离是事物之间差异性的度量，差异性越大，则想磁性越小，所以距离是系统聚类分析的依据和基础。常见的距离有：

（1）绝对距离

$$d_{ij}=\sum_{k=1}^{m}\left|x_{ik}-x_{jk}\right| \qquad (i,j=1,2,...,m) \qquad (4.2.5)$$

（2）欧氏距离

$$d_{ij}=\sqrt{\frac{1}{m}\sum_{k=1}^{m}(x_{ik}-x_{jk})^2} \qquad (i,j=1,2,...,m) \qquad (4.2.6)$$

（3）明科夫斯距离

$$d_{ij}=(\sum_{k=1}^{n}\left|x_{ik}-x_{jk}\right|^{p})^{\frac{1}{p}} \qquad (i,j=1,2,\ldots,m) \qquad (4.2.7)$$

式中：$p \geq 1$。当 $p=1$时，它就是绝对值距离；当 $p=2$ 时，它就是**欧氏距离**。

（4）切比雪夫距离

$$d_{ij} = \max_{k} \left| x_{ik} - x_{jk} \right| \quad (i, j = 1, 2, \ldots, m) \qquad (4.2.8)$$

选择不一样的距离，聚类结果会有所不同。在地理区分与分类研究中，往往采用几种距离进行计算，对比，选择一种较为合适的距离进行聚类。

据表4.2.3中的数据，用公式4.2.5，计算可得九个农业区之间的绝对值距离矩阵容下：

$$D=(d_{ij})_{9\times 9}=\begin{bmatrix} 0 & & & & & & & & \\ 1.52 & 0 & & & & & & & \\ 3.10 & 2.70 & 0 & & & & & & \\ 2.19 & 1.47 & 1.23 & 0 & & & & & \\ 5.86 & 6.02 & 3.64 & 4.77 & 0 & & & & \\ 4.72 & 4.46 & 1.86 & 2.99 & 1.78 & 0 & & & \\ 5.79 & 5.53 & 2.93 & 4.06 & 0.83 & 1.07 & 0 & & \\ 1.32 & 0.88 & 2.24 & 1.29 & 5.14 & 3.96 & 5.03 & 0 & \\ 2.62 & 1.66 & 1.20 & 0.51 & 4.84 & 3.06 & 3.32 & 1.40 & 0 \end{bmatrix}$$

三，直接聚类法

直接聚类法，是根据距离矩阵的结构一次并类得到结果，是一种简单的聚类方法。

例题：根据距离矩阵式（4.2.9），用直接聚类法对某地区的9个农业区进行聚类分析，步骤如下：

(1) 在距离矩阵D中，除去对角线元素以外，$d_{49}=d_{94}=0.51$为最小者，故将第4区与第9区并为一类，划去第9行和第9列；

(2) 在余下的元素中，除对角线元素以外，$d_{75}= d_{57}=0.83$为最小者，故将第5区与第7区并为一类，划掉第7行和第7列；

(3) 在第2步之后余下的元素之中，除对角线元素以外，$d_{82}= d_{28}=0.88$为最小者，故将第2区与第8区并为一类，划去第8行和第8列；

(4) 在第3步之后余下的元素中，除对角线元素以外，$d_{43}=d_{34}=1.23$为最小者，故将第3区与第4区并为一类，划去第4行和第4列，此时，第3，4，9区已归并为一类；

(5) 在第4步之后余下的元素中，除对角线元素以外，$d_{21}=d_{12}=1.52$为最小者，故将第1区与第2区并为一类，划去第2行和第2列，此时，第1，2，8区已归并为一类；

(6) 在第5步之后余下的元素中，除对角线元素以外，$d_{65}=d_{56}=1.78$为最小者，故将第5区与第6区并为一类，划去第6行和第6列，此时，第5，6，7区已归并为一类；

(7) 在第6步之后余下的元素中，除对角线元素以外，$d_{31}=d_{13}=3.10$为最小者，故将第1区与第3区并为一类，划去第3行和第3列，此时，第1，2，3，4，8，9区已归并为一类；

(8) 在第7步之后余下的元素中，除去对角线元素以外，只有$d_{51}=d_{15}=5.86$，故将第1区与第5区并为一类，划去第5行和第5列，此时，第1，2，3，4，5，6，7，8，9区均归并为一类。 根据上述步骤，可以作出聚类过程的谱系图4.2.1。

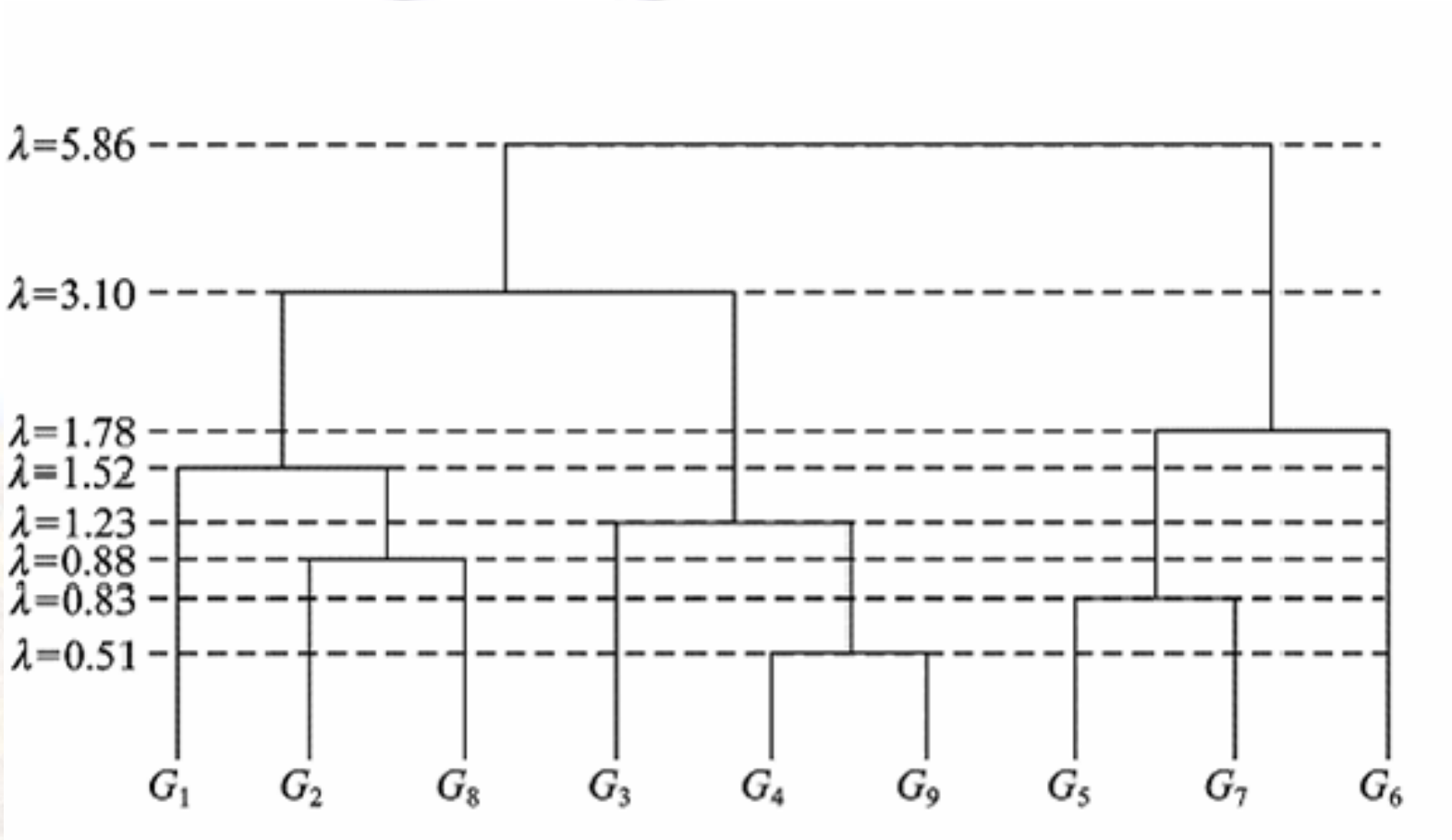

图 4.2.1 直接聚类谱系图

四，最短距离法

原理：最短聚类法，是在原来的 $m \times n$ 矩阵的 G_P 非对角元素中找出 $d_{pq} = \min\{d_{ij}\}$，把分类对象和 G_q 归并为一新 G_r 类，然后按计算公式：

$$d_{rk} = \min\{d_{pk}, d_{qk}\} \quad (k \neq p, q) \qquad (4.2.10)$$

实例分析：天山北坡绿洲城镇分类

绿洲城镇分类是指全面分析各城镇的自然、社会、经济方面的特征，并通过研究城镇形成的基本地理因素和综合因素，分类出绿洲经济比较明显的，类型特征相对属于一类的不同类型且相互间有明显差异的绿洲城镇。

根据表4.2.4提供的塔里木盆地34个绿洲型城镇17个分类指标平均值，标准化后的数据，下面我们运用系统聚类法，对该塔里木盆地34个绿洲型城镇进行聚类分析。

地理系统在数学方法和计算机信息技术的辅助下，形成和发展了许多复杂的聚类分析方法，目前的主要方法有：联接（树状聚类）分析法，逐步聚类分析法，双向联接聚类分析法等方法。此研究中采用了联接（树状聚类）分析法，验算了作为分类统计量的各种距离系数的算法，最后选用了CHEBYCHEV距离作为其分类统计量，其计算公式如下：

$$CHEBYCHEV(x,y)=\max|x_i-y_i| \quad (4.2.11)$$

式中，x_i 表示第一个样本在第 i 变量上的取值；y_i 表示第二个样本在第 i 变量上的取值。

天山北坡绿洲城镇分类

对于所选取的分析样本数据进行标准化处理以后，根据公式5.2.11进行计算，做出的聚类分析谱系图（此过程可以在SPSS、STATICTICA或MATLAB软件的支持下完成）。

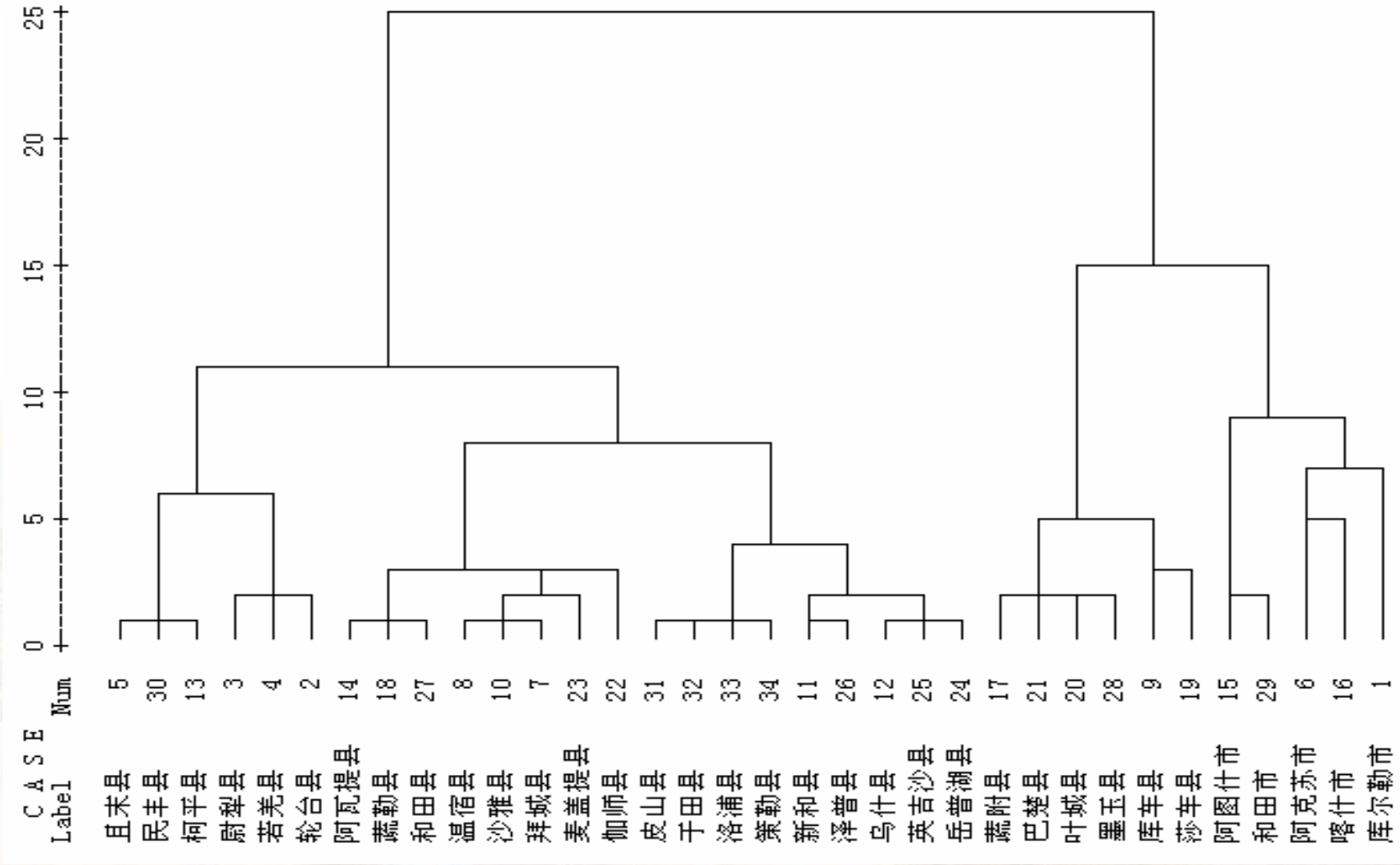

图4.2.4 塔里木盆地绿洲城镇的聚类分析谱系图（1990～1994年）

根据实际城镇各因素间的相似程度和相关程度把塔里木盆地绿洲城镇共分为以下三类。

表4.2.7 1991～1995年塔里木盆地34个绿洲城镇分类表

类别	第一类		第二类			第三类		
城镇名称	库尔勒市		拜城县	疏勒县	墨玉县	轮胎	麦盖提县	洛浦县
	阿克苏市		温宿县	乌什县		尉犁县	岳普湖县	策勒县
	喀什市		库车县	叶城县		若羌县	英吉沙县	于田县
	阿图什市		莎车县	巴楚县		且末县	泽普县	和田县
	和田市		阿瓦提县	伽师县		新和县	民丰县	
			疏附县	沙雅县		柯坪县	皮山县	

Ward 方法
CHEBYCHEV 距离

图4.2.5 塔里木盆地绿洲城镇的聚类分析谱系图（1995～1999年）

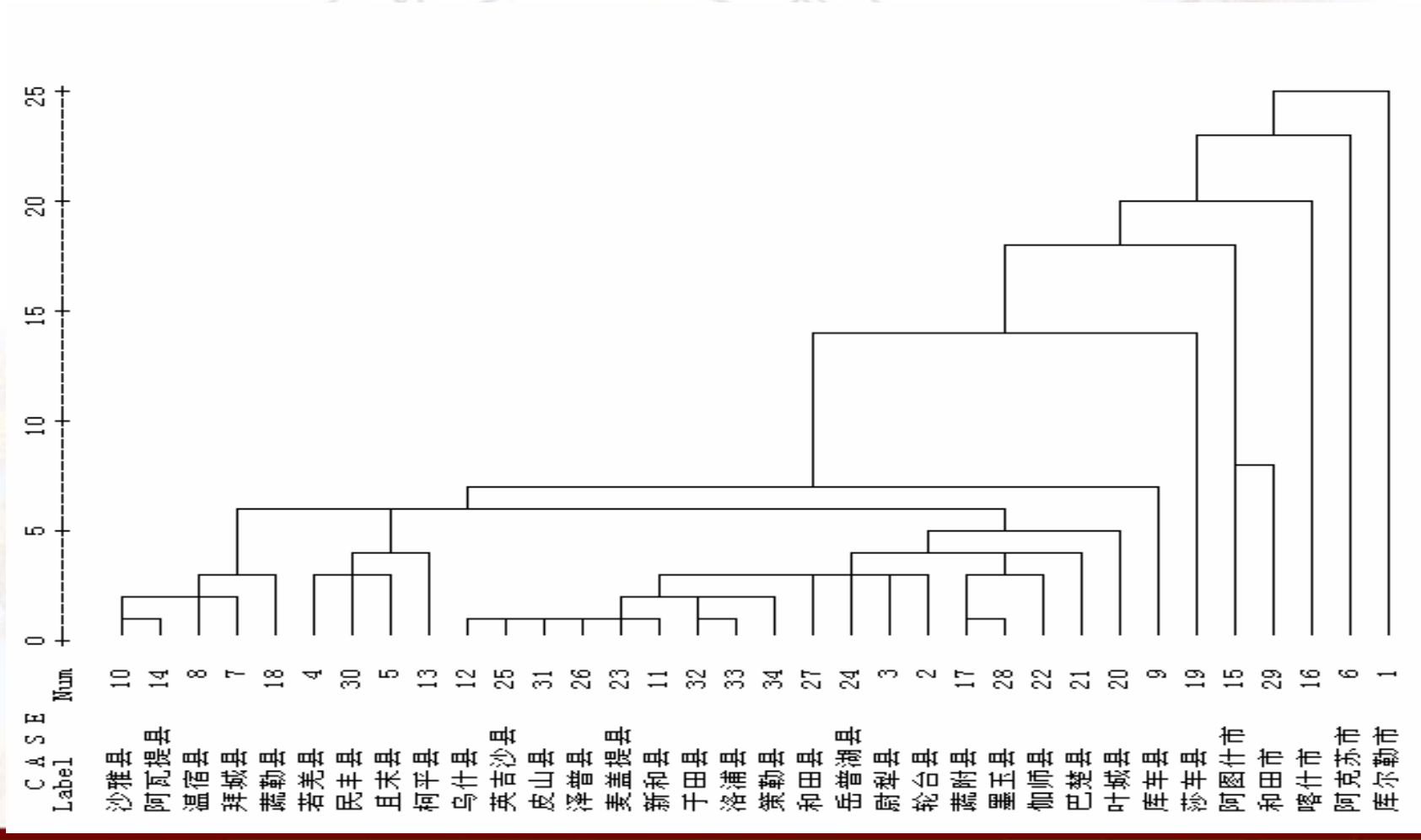

表4.2.9 2001-2006年塔里木盆地34个绿洲城镇分类表

类别	第一类		第二类			第三类		
城镇名称	阿克苏市		库车县	蔬勒县	墨玉县	尉犁县	麦盖提县	洛浦县
	库尔勒市		莎车县	轮台县	和田县	乌什县	岳普湖县	策勒县
	喀什市		拜城县	叶城县		若羌县	英吉沙县	于田县
	和田市		温宿县	巴楚县		且末县	泽普县	
	阿图什市		阿瓦提县	伽师县		新和县	民丰县	
			蔬附县	沙雅县		柯平县	皮山县	

分类结果分析：

第一类：区域经济中心城镇　通过对1990年到2005年的统计数据进行聚类分析得出库尔勒市、阿克苏市、喀什市、和田市和阿图什市等这类绿洲城镇都是各地洲的经济、政治、文化和科技中心。统计资料显示，这类城镇在所在区域国内生产总值占有举足轻重的地位。

第二类：区域资源主导型城镇　通过聚类分析得出，塔里木盆地绿洲城镇中除了第一类的城镇类型之外还有些城镇在拥有自然资源方面占优势。该类的城镇是自古以来就是塔里木盆地重要的农牧业大县。其中莎车县、叶城县、巴楚县和伽师县等县离中心城市远，工业基础薄弱，但其农业产值在当地国内生产总值中占相当大的一部分。

第三类：区域经济潜力型城镇　通过聚类分析得出，这类绿洲城镇应属于区域经济潜力型城镇。该类的城镇在经济发展状况，资源位置优势方面远低于上述两类。其中且末县、若羌县、民丰县和尉犁县等，各自间优势项目差异较大，且末县和若羌县自然资源比较丰富；民丰县、尉犁县和和田县位于区域交通枢纽中心地带，畜牧业所占比例较大，荒漠草原与待开发荒地开发潜力较大，所以，区域发展中应以规模性农业和农区畜牧业为主要方向，加大石油天然气和其它矿物资源的开发力度，因此应加快发展交通运输服务业以及种植业，畜牧业等农副产品的深加工。麦盖提县、岳普湖县、英吉沙县、策勒县、于田县和皮山县等离区域中心城市较远，自然条件较差，生态环境脆弱等不利因素，所以这些县大多数属于贫困县。在区域发展中，应以生态农业为基础，发挥各自优势，扩建县级市场，加快发展旅游业和地方产业，吸纳人才，把该类城镇建设成中型城市作为其发展的方向。