

第七章 主成分分析、因子分析、聚类分析及其在城市社会区分析中的应用

127

本章介绍三种重要的多元统计分析方法：主成分分析（PCA）、因子分析（FA）和聚类分析（CA）。主成分分析和因子分析常常一起用于约减变量数，在消除变量共线性及揭示潜在变量方面尤其有用，二者在社会经济研究中具有广泛的应用（也可参见第 8.4 节的案例 8）。主成分分析和因子分析是将变量合并成组，聚类分析是将观察样本按属性的相似性进行分类。换言之，给定一个数据表，主成分分析和因子分析用于减少数据的列数（变量数），聚类分析减少其行数（样本数）。

本章以城市社会区分析的例子来演示上述三种方法的应用。在解释分析结果时，我们比较了城市结构的三种经典模型：同心圆模型、扇形模型和多核心模型。下面的分析也演示了如何用严谨的数量分析法将各种描述性的模型整合在一个框架下。本章以北京为例应用上述三种方法，并用 GIS 绘制空间模式图。

第 7.1 节介绍主成分分析和因子分析。第 7.2 节介绍聚类分析。第 7.3 节简单介绍相关社会区分析。第 7.4 节以北京社会空间结构分析为例，提供了中国城市结构快速转变中的一个新视角。第 7.5 节是讨论和小结。

7.1 主成分分析和因子分析

主成分分析和因子分析常用于变量约减。其作用主要表现在两方面，一是借此发现一些潜变量，从而简化和事物的结构、为分析研究对象带来方便；二是消除变量之间的多重共线性，为后续的回归分析服务。在许多社会经济应用领域，从数据中提取的原始变量常常彼此相关，包含一定程度的重复信息。主成分分析和因子分析对原始变量进行约减，从而简化了分析结构。得到的主成分或因子之间相互独立、不再相关（假设不进行旋转或用正交旋转），可以作为回 128

归分析的解釋变量。

尽管主成分分析和因子分析有很多共同之处，但二者“在概念和数学基础方面都非常不一样”（Bailey and Gatrell, 1995: 225）。主成分分析只是简单地对原始数据进行变换，变量（成分）个数不变，因此它是一个数学变换（严格地讲，并不是一个统计运算）。因子分析以较少

的变量（因子）来承载原始变量的大多数信息（当然有一些误差项），因而是一个实实在在的统计分析过程。主成分分析试图解释各观测变量的方差，而因子分析旨在解释观测变量之间的相关性（Hamilton, 1992: 252）。在许多应用中（包括本书的案例），两种方法常常结合起来使用。在 SAS 中，主成分分析是因子分析下面的一个备选子操作。

7.1.1 主成分因子模型

主成分分析（PCA）的原理是将 K 个原始观测变量 Z_k 变换为 K 个彼此独立（互不相关）的主成分 F_k ：

$$Z_k = l_{k1}F_1 + l_{k2}F_2 + \dots + l_{kj}F_j + \dots + l_{kK}F_K \quad (7.1)$$

当只保留最大的 J 个成分时 ($J < K$)，有

$$Z_k = l_{k1}F_1 + l_{k2}F_2 + \dots + l_{kJ}F_J + v_k \quad (7.2)$$

其中，被舍弃的成分归入残差项 v_k 中，即

$$v_k = l_{k,J+1}F_{J+1} + l_{k,J+2}F_{J+2} + \dots + l_{kK}F_K \quad (7.3)$$

式 7.2 和 7.3 为主成分因子分析（PCFA）模型，它保留承载大部分信息的若干主成分，舍弃了包含信息少的次要成分。社会区分析中用的就是 PCFA 这种方法（Cadwallader, 1996: 137），在本章余下部分将其简单地称为“因子分析”，但这并不是真正的因子分析。

在真正的因子分析（FA）中，残差项为 u_k ，与 PCFA 中的 v_k 不同的是，每个 Z_k 变量的残差项不同：

$$Z_k = l_{k1}F_1 + l_{k2}F_2 + \dots + l_{kJ}F_J + u_k$$

这里的 u_k 称为特殊因子（与普通因子 F_j 相对）。在 PCFA 中，残差项 v_k 是被舍弃变量 (F_{J+1}, \dots, F_K) 的线性组合，因而不可能像真正因子分析中的 u_k 那样彼此不相关（Hamilton, 1992: 252）。

7.1.2 因子载荷、因子得分和特征值

为方便起见，在进行主成分和因子分析之前先要对变量 Z_k 的原始观测值标准化；主成分

(因子)的初始值也标准化了。数据标准化即将一系列数据 x 转换为新的数据 x' , 转换后的数据平均值为 0, 标准差为 1: $x' = (x - \bar{x}) / \sigma$ 。当 Z_k 和 F_j 都标准化后, 式 7.1 和 7.2 中的 l_{kj} 称为变量 Z_k 与主成分 (因子) F_j 之间回归的标准化系数, 又称为因子载荷。例如, l_{kl} 是变量 Z_k 在主成分 F_l 上的载荷。因子载荷反映了变量与因子之间关系的强弱。

反过来, 主成分 F_j 也可以表作为原始变量 Z_k 的线性组合:

$$F_j = a_{1j}Z_1 + a_{2j}Z_2 + \dots + a_{Kj}Z_K. \quad (7.4)$$

这些主成分 (因子) 的估计值称为因子得分。 a_{kj} 称为因子得分系数, 即因子与变量之间的回归系数。

主成分 F_j 彼此不相关, 其排列顺序为, 第一个主成分 F_1 具有最大的样本方差 (λ_1), 主成分 F_2 有第二大方差, 依此类推。与主成分对应的方差 λ_j 称为特征值, 即有 $\lambda_1 > \lambda_2 > \dots$

因为标准化变量的方差为 1, 从而所有变量的方差之和等于变量数, 即

$$\lambda_1 + \lambda_2 + \dots + \lambda_K = K \quad (7.5)$$

因此, 第 j 个主成分解释的方差比例为 λ_j/K 。

根据特征值可以判断主成分 (因子) 的重要性, 从而决定选择多少个主成分。比如, 我们可以定义特征值大于 1 的为重要主成分 (Griffith and Amrhein, 1997: 169)。因为标准化变量的方差为 1, 那么任何特征值 $\lambda < 1$ 的主成分都比原始变量的方差还小, 也就是说这个主成分抓住的信息还不如原来变量包含的信息量大, 从而没有起到变量约减的作用。

选特征值大于 1 为主成分的标准是主观的。实际操作时可以参考特征值与主成分 (因子) 之间的碎石图 (Hamilton, 1992: 258)。例如, 图 7.1 为 14 个主成分的特征值碎石图 (来自 7.4 的案例 7)。由图可知, 在第 4 个主成分之后发生明显转折, 碎石图趋于平缓, 表明第 5~14 个主成分解释的方差相对较小。因此, 可以保留 4 个主成分。

SAS 等统计分析软件的输出结果中包括因子载荷、特征值和解释总方差比例等重要信息。因子得分可以保存为事先指定名称的文件。SAS 的因子分析还可以得到一个原始观测变量的相关矩阵用于检验他们的相关性。

7.1.3 旋转操作

往往变量载荷平均分散于多个因子中，解释主成分因子分析得到的原始结果有一定困难。旋转操作并不影响对数据的拟合程度，但可以使变量在某一个因子上载荷最大（或正或负），而在其他因子上的载荷较小，从而简化因子结构，从而使得到的因子便于解释。最后，我们可以根据因子（潜变量）中载荷大的变量组合情况命名各因子。

正交旋转得到独立（彼此不相关）的因子，这一点很重要。方差极大旋转法是一种广泛使用的正交旋转法，它使每个因子载荷平方的方差最大，从而极化因子载荷（即变量在某个因子上的载荷或者很高、或者很低）。社会研究中常常使用方差极大旋转法。斜交旋转（例如最大斜交旋转）的极化效果更大，但是允许因子间存在一定程度的相关。在 SAS 中，提供了多种旋转方式供选择。

图 7.2 是主成分因子分析的小结：

1. 原始数据的标准化：所得结果（ Z ），变量个数（ K ）和观测记录数（ n ）保持不变。
2. 主成分分析（PCA）：用 K 个彼此不相关的主成分来解释 K 个变量的全部方差。
3. 主成分因子分析（PCFA）：用 J （ $J < K$ ）个主成解释大部分方差。
4. 因子旋转：使每个变量在某一个因子上的载荷最大（在其他因子上的载荷很小，甚至接近于 0），以增加解释能力。

SAS 软件中的因子分析（FA）为 FACTOR 操作，可以同时得到因子分析之前的主成分分析（PCA）结果。下面是一个因子分析的 SAS 示例语句，用 4 个主因子来解释 14 个原始变量（x1 到 x14），使用了方差极大旋转法。

```
proc factor out=FACTSCORE (replace=yes)
    nfact=4 rotate=varimax;
    var x1-x14;
```

131

上面语句中的 FACTSCORE 为因子得分，可以另存为外部文件。应该提到的是，SAS 程序中语句是大写还是小写并无关系。

7.2 聚类分析

聚类分析（CA）是把观测样本根据相似性进行分组。根据聚类标准，分组后组内样本的相似性比组间样本的相似性大。需要注意的是，聚类分析跟另外一种类似的多元统计分析——判别分析（DFA）之间的区别。二者都是根据特征变量将观测样本进行分类，聚类分析是根据观测样本确定类别（类别事先并不知道），而 DFA 是先给定类别再对判断观测样本究竟属于那一类（即事先知道类别）。

地理学家一直对聚类分析很感兴趣，将其广泛应用于区划和城市分类等领域。在社会区分析研究中，聚类分析用在对因子分析的结果（即主因子得分不同的各地区）进行分类，从而得到不同类型的社会区。

对观测数据聚类时的一个重要指标是“属性距离”，它可以有各种不同的测量方法。最常用的是欧式距离：

$$d_{ij} = \left(\sum_{k=1}^K (x_{ik} - x_{jk})^2 \right)^{1/2}, \quad (7.6) \quad 132$$

这里， x_{ik} 和 x_{jk} 是 K 维对象 i 和 j 的第 k 个变量值。当 $K = 2$ 时，欧式距离即为二维平面上 i 和 j 的直线距离。像第二章讨论的那样，这里的距离也包括曼哈顿距离、明可斯克(Minkowski)距离、坎倍拉(Canberra)距离等（Everitt et al., 2001: 40）。

层次聚类法大致分为两种，即自下而上的凝聚式（Agglomerative）和自上而下的分裂式（Divisive）。凝聚式方法开始将每个观测数据分为一类（也称为“簇”），然后每次寻找最近的两个类进行合并，直到所有观测数据合并为一个类。分裂式方法开始将所有观测数据视为一个类，然后每次选择最大的一个类分裂为两个，直到每个观测数据自成一类。这里重点介绍使用最广的凝聚式层次聚类法（AHMs）。按这种规则分类得到的结果可以归结为一个树状图，显示了逐级分类的过程。图 7.3 是下面例子的分类结果。在树状图中，分类类型逐级嵌套，每一类都是一个更大更高级类的一个元素。

下面是层次聚类法的一个示例，用了单链法或称最短距离法。数据库中一共有 4 个观测数

据，其距离矩阵如下：

$$D_1 = \begin{matrix} & \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 6 & 5 & 0 & \\ 9 & 7 & 4 & 0 \end{bmatrix} \end{matrix}$$

上述矩阵中的最小非零项 D_1 为 $(2 \rightarrow 1) = 3$ ，因此 1 和 2 最先分为一类即 C1。根据最短距离 133

法则，C1 与其他数据之间的距离为：

$$d_{(12)3} = \min\{d_{13}, d_{23}\} = d_{23} = 5$$

$$d_{(12)4} = \min\{d_{14}, d_{24}\} = d_{24} = 7$$

于是得到一个新的矩阵，它由 C1、3 和 4 两两之间的距离组成：

$$D_2 = \begin{matrix} & \begin{matrix} (12) \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} 0 & & \\ 5 & 0 & \\ 7 & 4 & 0 \end{bmatrix} \end{matrix}$$

这时最小的非零项 D_2 为 $(4 \rightarrow 3) = 4$ ，从而 3 和 4 分为一类即 C2。最后，C1 和 C2 的距离为 5，二者同属于 C3 类，C3 包含了全部 4 个观测数据。分类过程可用图 7.3 的树状图表示，其纵坐标高度代表每一次组合的距离。

类似的，完全联系法（最长距离法）用两个观测数据（分别属于两类）之间的最长距离来进行分类；平均距离法使用两个观测数据之间的平均距离；重心法使用观测数据与类平均值（重心）之间的欧式距离平方。

另外一个常用的层次聚类法是沃德法（Ward's method），其目的是从总体上使每一步的组内误差平方和最小，误差平方和表作

$$E = \sum_{c=1}^C E_c$$

这里

$$E_c = \sum_i^{n_c} \sum_{k=1}^K (x_{ck,i} - \overline{x_{ck}})^2$$

其中 $x_{ck,i}$ 是第 c 类中变量 k 的第 i 个观测值, \bar{x}_{ck} 为 c 类中变量 k 的平均值。

每种聚类方法各有优缺点。一个合适的聚类应该是各类的规模相近, 分布相对集中, 形态紧凑、内部均质 (Griffith and Amrhein, 1997: 217)。单链法得到的结果一般是非均匀零散聚类 134, 应尽量避免。如果偏远样本是需要考虑的主要因素, 应该用质心法。如果要得到紧凑的聚类, 应该用最长距离法。沃德法的结果为规模相当的球形聚类, 如没有需要特别考虑的因素可选用这种聚类法 (Griffith and Amrhein, 1997: 220)。本章的案例就用了沃德法。

聚类个数根据特定的应用目的而定。与因子分析中根据特征值确定主因子的方法类似, 我们也可以用碎石图来确定。在沃德法中, 可以用对应的 R^2 与聚类个数的碎石图来确定。越过某一聚类数后进一步融合类别并不能增加多少均质性。

在 SAS 中, 用 CLUSTER 命令来实现聚类分析, 用 TREE 命令可以生成树状图。下面的 SAS 语句使用了沃德法, 聚类树状图中最多包含 9 个小类。

```
proc cluster method=ward outtree=tree;

    id subdist_id; /* variable for labeling ids */

    var factor1-factor4; /* variables used */

proc tree out=bjcluster ncl=9;

    id subdist_id;
```

7.3 社会区分析

社会区分析发轫于希维克和威廉姆斯 (Shevky and Williams, 1949) 对洛杉矶的住宅区分异性研究, 此后, 希维克和贝尔 (Shevky and Bell, 1955) 在对旧金山的同类研究中做了进一步发展。研究的基本问题是由社会分化导致的城市居住空间分异。根据经济地位 (社会等级)、家庭结构 (城市化) 和种族情况 (种族隔离) 三个基本要素将普查小区分成不同类型的社会区。最初的研究是从六个要素简化成上述三个要素: 经济地位对应于原来的职业和教育; 家庭结构对应于家庭人口数、妇女就业状况及单亲家庭; 种族状态对应于少数民族百分比 (Cadwallader, 1996: 135)。在因子分析中, 理想的因子载荷可能象表 7.1 的样子。后来的研究

使用了更多的变量，进一步验证了上面提取的三个要素的有效性（Berry, 1972: 285; Hartshorn, 1992: 235）。

地理学家通过分析上述要素的空间分布模式，在社会区分析中取得了重要进展（如 Rees, 1970; Knox, 1987）。社会经济地位因子往往表现出一种扇形模式：高收入和高教育阶层汇集成一个或几个扇形区域，而低收入和低教育汇集到其他的扇形区域。家庭结构因子往往呈同心圆分布：里层是年轻人或老年人组成的小家庭，外层是中年人组成的大家庭。种族因子则以特定的种族为中心形成一个个聚集区。三种要素共同作用下形成一种复杂的“城市万花筒”(urban mosaic)，通过聚类分析可以得到不同类型的社会区域，如图 7.4 所示。通过城市社会区的分析，我们把布杰斯的同心圆模型（Burgess, 1925）、侯以德的（Hoyt, 1939）扇形模型和乌尔曼-哈里斯多核心模型（Harris and Ullman, 1945）整合在同一个框架之内。换言之，这三种模型从不同的角度反映了城市的复杂结构，彼此互补。

对用上述因子生态法解释城市居住空间分异的批评至少有三种（Cadwallader, 1996: 151）。首先，分析结果很容易受到研究变量、分析单元、因子分析方法等因素的影响。第二，它仍然是一种描述性分析法，不能解释产生这种模式的深层原因。第三，这种方法确定的社会区域是同质的，但并不一定就是功能性区域或联系紧密的社区。尽管存在这些批评，社会区分析有助于我们理解城市的居住分异，是一种研究城市内部社会空间结构的重要工具。社会区分析在发达国家尤其是北美城市研究中得到广泛应用（相关研究请参见 Davies and Herbert, 1993），在发展中国家的应用也不少（参见 Berry and Rees, 1969; Abu-Lughod, 1969）。

7.4 案例 7：北京的社会区分析

下面的案例是基于顾朝林等人的一项研究提取设计的（Gu et al., 2005），详细的研究过程和结果解释可以阅读原文。本节重点演示如何实现三种统计方法。此外，我们还演示了用虚拟变量回归模型来检验各因子得分的空间结构模式。自 1978 年经济改革，尤其是 1984 年城市改革（包括城市土地利用改革和住房改革等方面）以来，中国的城市景观发生了重大变化。许多大城市逐渐从自给自足的工作单位邻域系统向更加多元化的城市空间转变。作为中国的首都，

北京为我们提供了中国城市结构转变的绝佳案例。

研究区域为北京城八区的连续城市化地区，包括 107 个乡镇、街道，不包括外围的两个区（门头沟和房山）以及城八区边缘的 23 个乡镇（主要为乡村地区，也缺乏完整的数据），如图 7.5 所示。1998 年，研究区内的总人口为 590 万人，平均每个乡镇街道人口为 55200 人。数十年来，乡镇街道一直是北京的基本行政管辖单位，也是我们能从政府那里获取统计数据的最小单元，因而是我们研究的分析单元。由于在全国性的人口普查中缺乏社会经济统计数据，本例的主要数据来自 1998 年北京市分区统计年鉴。有些数据如个人收入和居住空间情况来自 1998 年的住户调查。

本例所需数据如下：

1. 包括 107 个乡镇街道边界的 **shape** 文件 `bjsa`;
2. 各乡镇街道的属性数据文件 `bjattr.csv`。

在 **shape** 文件 `bjsa` 中，将城区分为四个象限，用变量名 `sector` 表示（1 为东北，2 为东南，3 为西南，4 为西北）。城区从内向外分为 4 个环带，用变量名 `ring` 表示（1 为最里面的区域，2 为向外紧接的一个区域，如此类推）。象限和环带的划分主要用于分析社会空间结构。**Shape** 文件 `bjsa` 和属性数据文件 `bjattr.csv` 都包含一列 `ref_id`（乡镇街道编号），用于链接二者的共同列。属性文件 `bjattr.csv` 有 14 个社会经济变量（**X1-X14**），变量名及其基本的统计信息见表 7.2。

1. 用 **SAS** 进行主成分分析：参见附录 7B 的 **SAS** 程序 `FA_Clust.sas`（光盘提供）。第

一部分调用 **SAS** 里面的 **PROC FACTOR** 来进行主成分因子分析（**PCFA**）。程序调用属性数据文件 `bjattr.csv`，用 4 个因子来提取 14 个原始变量（`x1, x2, ..., x14`）的大部分信息。所得因子得分数据保存在 `factscore.csv` 文件中，包括 14 个原始变量及 4 个因子得分。

SAS 里面的 **FACTOR** 操作也输出了因子分析（**FA**）之前的主成分分析（**PCA**）结果。因子的个数（这里为 4 个）是通过考察主成分分析的特征值得到的（见表 7.3）。因

因子个数的选择不会影响主成分分析的结果。如果我们随意地确定一些因子个数（如 3 或 5），得到的主成分分析结果是一样的。为了确定因子分析中究竟要用多少个主成分（因子），需要考虑两个因素：包含的主成分越多，因子解释原始数据总方差的百分比越高；但主成分越多，因子的抽象性越差，结果越难解释。这就要求我们从两者中找到一个合适的平衡点。根据特征值大于 1 的标准（见第 7.1.2 节），我们保留了 4 个因子，大约解释了总方差的 70%。这从图 7.1 所示的碎石图看来也是合适的。

我们用方差最大旋转法进行因子旋转，这样变量在某个因子上的载荷最大，而在其他 139 因子上的载荷最小。表 7.4 列出了旋转后的因子结构（重新排列了变量顺序以便突出因子载荷结构）。四个因子反映了主载荷变量的信息，其命名如下：

- a. “土地利用强度”：最重要的变量，解释了总方差的 35.16%，主要包含 6 个变 140 量的信息，即 3 个密度变量（人口密度、公共服务设施密度、办公业和零售业密度）、2 个人口统计变量（就业率、抚养比）及住房价格。
- b. “邻里变量”：解释了总方差的 15.42%，包含 3 个变量，即流动人口比重、家庭人口数、人均住房面积。
- c. “社会经济地位”：解释了总方差的 10.42%，包含 2 个变量，即年人均收入水平、人口自然增长率。
- d. “种族”：解释了总方差的 9.22%，包含 3 个变量，即少数民族聚居区、性别比例、工业密度。

2. 用 SAS 进行聚类分析：上述 SAS 程序 FA_Clust.sas 第二部分调用 PROC CLUSTER 进行聚类分析，得到一个聚类树状图。在具体操作时，确定的聚类数目 NCL=5，此即为聚类树图的下限。聚类结果保存在文件 cluster5.csv 中（将列 cluster 改名为 cluster5 以示区别）。设置聚类数 NCL=9，重新进行聚类分析，结果保存为文件 cluster9.csv（将列 cluster 改名为 cluster9）。

这里先聚为 5 类，后又扩展到 9 类，从而可以揭示更细的空间分布态势的信息。例如

，当分为 9 类时，原来聚为 5 类时的第 2 类被再分为 2、4、5 三类，每一类代表一种社会区。

3. 用 ArcGIS 绘制因子分布图：用 ArcGIS 打开 shape 文件 bjsa，根据 ref_id 连接属性文件 factscore.csv，绘制因子得分图。图 7.6a 为 factor1（土地利用强度），图 7.6b 为 factor2（邻里变量），7.6c 为 factor3（社会经济地位），图 7.6d 为 factor4（种族）。

4. 用 ArcGIS 绘制社会区域图：类似第 3 步，在 ArcGIS 中将 cluster9.csv 和 cluster5.csv 链接到 shape 文件 bjsa，绘制社会区域图，见图 7.7。5 种基本的社会区域用不同的图形模式表示，9 种更详细的类型以类编号标志。

为了理解每种社会区域的特点，可以用 ArcGIS 里的“summarize”工具对融合后的属性表进行处理，得到每类的因子平均得分，结果见表 7.5。各类按因子得分即距离城市中心的距离进行命名。

5. 用虚拟变量回归以考察因子空间结构：借助回归模型，可以考察因子空间分布态势是否具有一定的结构特征（如这里的环状或扇形等）（Cadmwallader, 1981）。北京基于环形道路骨架，可以分为 4 个环形区域，用 3 个虚拟变量表示（ x_2 , x_3 和 x_4 ）。类似的，用另外 3 个虚拟变量（ y_2 , y_3 和 y_4 ）代表 4 个扇形区域（NE, SE, SW 和 NW）。表 7.6 列出了虚拟变量与环形区域和扇形区域的对应关系。

衡量环形空间结构的一个简单线性模型可以表作

$$F_i = b_1 + b_2x_2 + b_3x_3 + b_4x_4, \quad (7.7)$$

其中， F_i 为各个街道乡镇的因子得分（有四个因子 $i = 1, 2, 3$ 和 4），常数 b_1 为环带 1（ $x_2=x_3=x_4=0$ ，也称为参考环带）的平均因子得分，系数 b_2 、 b_3 、 b_4 分别为环带 2、3、4 与环带 1 因子得分之间的平均差值。类似的，可以用下述模型考察扇形空间结构

$$F_i = c_1 + c_2y_2 + c_3y_3 + c_4y_4, \quad (7.8)$$

这里各变量和常数的意义与式 7.7 类似。

将 shape 文件 bjsa（与 factscore.csv 链接后）的属性表输出为外部文件 zone_sect.dbf，里面包含因子得分、环带编号 ring 和扇形编号 sector。在文件 zone_sect.dbf 基础上，用 Excel 或 SAS 基于表 7.6 创建并计算虚拟变量 x_2 , x_3 , x_4 , y_2 , y_3 和 y_4 ，然后根据式 7.7 和 7.8 进行回归分析。回归结果见表 7.7。本书光盘提供了这里所需的 SAS 程序 BJreg.sas。

7.5 讨论与结论

表 7.7 中的 R^2 显示了环带或扇形模型的回归效果，括号内的 t 值揭示了系数的统计显著性（即某个环带或扇形区域是否与参考环带或参考扇形的因子得分显著不同）。显然，土地利用强度很近似于环形模式，负系数 b_2 , b_3 和 b_4 都通过显著性检验，表明土地利用强度从中心向外衰减。邻里因子呈现出较好的扇形分布，正系数 c_4 （通过显著性检验）表明北京西北地区流动人口比重较高。社会经济地位因子同时具有环形和扇形的分布态势，但扇形分布趋势更强。负系数 b_3 和 b_4 （都通过显著性检验）表明因子得分在第三、四个区域有所降低；而正系数 c_3 （144 通过显著性检验）表明西南扇形的因子得分较高，因为这里包含宣武区的两个高收入街道。种族因子既不呈环带分布，也不呈扇形分布。种族聚居点散布于全城，可能用一种多中心模型表示更合适。

显然，土地利用强度是影响北京同心圆社会空间结构的主要因素。从城市中心区（第 4、6、8、9 类）到近郊区（第 1、2 类）再到远郊区（第 3、5、7 类），人口密度、公共服务设施密度、办公和零售业密度随土地价格下降而衰减。主要受流动人口影响的邻里变量，是影响北京社会区域的第二个因子。大量外来人口聚居在增长快速、经济机会多的海淀区（第 1 类）以及制造业岗位多的石景山区（第 3 类）。第三个因子（社会经济地位）的作用主要表现在两方面：一是两个内城街道（第 8 类）高收入地区的浮现，二是中等收入（第 1 类）和低收入（第 2、3、5 类）之间的分异。只有当类型扩展到 9 类时，第四个因子（种族）的作用才显现出来。

在西方，社会经济地位是形成城市扇形模式的一个主导力量，家庭结构促成城市的环形结构，种族分布显示一种多中心态势。对北京市而言，社会经济地位和种族因子依然有作用，但其影响不如西方城市重要，而家庭结构因子在北京基本上不起作用。在发达国家，大部分城市的人口普查数据及对应的空间数据（例如美国的 TIGER 数据）获取十分方便，因而对那些城市进行社会区分析非常容易。但是，对于发展中国家，可靠数据来源常常成为城市社会区域研究的一个巨大障碍。随着数据质量的提高，例如包含的社会经济、人口统计和住户等变量数越来越多，统计单元越来越小，对未来开展这类研究创造了条件。

附录 7A 判别分析

事物按特征可以分成不同的类型，并可以用数量方法来描述。判别分析（DFA）的主要目的是根据那些描述事物特征的变量建立一个线性方程，借此将观察对象归入已知的类型中。DFA 与聚类分析不同，后者的类型事先并不知道。例如，男女骨头的结构不同，但我们已知性别只有两类。当我们发现一些遗骨时，就可以用 DFA 法来确定其性别。

下面是 DFA 的一个示例，所用对象由两个类别组成。例如，我们有两类物体，A 和 B，用变量 p 测量。第一类有 m 个观测值，第二类有 n 个观测值，从而观测值记为

$$X_{ijA} (i = 1, 2, \dots, m; j = 1, 2, \dots, p)$$

$$X_{ijB} (i = 1, 2, \dots, n; j = 1, 2, \dots, p).$$

我们的目标是寻找下述判别方程 R

$$R = \sum_{k=1}^p c_k X_k - R_0 \quad (\text{A7.1})$$

其中， c_k ($k = 1, 2, \dots, p$) 和 R_0 为常数。

将 m 个观测值 X_{ijA} 代入 R 后，得到 m 个 $R(A)$ 值。类似的，可以得到 n 个 $R(B)$ 值。 $R(A)$ 和 $R(B)$ 值都满足一定的统计分布。我们需要找到一个方程 R ，使得 $R(A)$ 和 $R(B)$ 的分布彼此分离很远。为了达到这个目的，需要下面两个条件

1. 平均值之间的差距 $Q = \overline{R(A)} - \overline{R(B)}$ 最大；

2. 方差之和 $F = S_A^2 + S_B^2$ 最小（即数据沿曲线在峰值附近呈紧凑分布）。

上述条件等价于寻找系数 c_k 使得 $V = Q/F$ 最小。一旦得到系数值 c_k ，就可以简单地用 146

$R(A)$ 和 $R(B)$ 的平均值代替 R_0 :

$$R_0 = [\overline{mR(A)} + \overline{nR(B)}]/(m+n) \quad (A7.2)$$

对于给定的样本，先计算 R 值，再与 R_0 进行比较。如果它比 R_0 大，则属于 A 类，否则属于 B 类。

DFA 可以调用 SAS 里面的 PROC DISCRIM、PROC STEPDISC 或 PROC CANDISC 等来实现。

附录 7B 因子分析和聚类分析的示例程序

```
/*FA_Clust.SAS runs Factor Analysis & Cluster Analysis
   for social area analysis in Beijing */
/*By Fahui Wang on 2-4-05 */

/* read the attribute data */
proc import datafile="c:\gis_quant_book\projects\bj\bjattr.csv"
  out=bj1 dbms=dlm replace;
  delimiter=', ';
  getnames=yes;
proc means;

/* Run factor analysis */
proc factor out=fscore(replace=yes)
  nfact=4 rotate=varimax; /* 4 factors used */
  var x1-x14;
/*export factor score data */
proc export data=fscore dbms=csv
  outfile="c:\gis_quant_book\projects\bj\factscore.csv";

/* Run cluster analysis */
/* Factor scores are first weighted by relative importance
   i.e., variance portions accounted for (based on FA) */
data clust; set fscore;
  factor1 = 0.3516*factor1;
  factor2 = 0.1542*factor2;
  factor3 = 0.1057*factor3;
  factor4 = 0.0922*factor4;
proc cluster method=ward outtree=tree;
  id ref_id; var factor1-factor4; /*plot dendrogram */
proc tree out=bjclus ncl=9; /*cut the tree at 9 clusters*/
```

```
    id ref_id;  
/* export the cluster analysis result */  
proc export data=bjclus dbms=csv  
    outfile="c:\gis_quant_book\projects\bj\cluster9.csv";  
run;
```

表 7.1 社会区分析的理想因子载荷

	经济地位	家庭状况	种族情况
职业	I	O	O
教育	I	O	O
人口数	O	I	O
妇女就业状况	O	I	O
单亲家庭	O	I	O
少数民族百分比	O	O	I

注：I 表示接近 1 或-1 的数；O 表示接近 0 的数。

表 7.2 北京市社会经济结构的基本统计参数 (n=107)

序号	变量	均值	标准差	最小值	最大值
X1	人口密度 (人/km ²) ¹	14,797.09	13,692.93	245.86	56,378.00
X2	自然增长率 (‰)	-1.11	2.79	-16.41	8.58
X3	性别比 (M/F)	1.03	0.08	0.72	1.32
X4	就业率 (%) ²	0.60	0.06	0.47	0.73
X5	家庭规模 (人/户)	2.98	0.53	2.02	6.55
X6	抚养比 ³	1.53	0.22	1.34	2.14
X7	收入 (元/人)	29,446.49	127,223.03	7,505.00	984,566.00
X8	公共服务设施密度 (个/km ²) ⁴	8.35	8.60	0.05	29.38
X9	工厂密度 (个/km ²)	1.66	1.81	0.00	10.71
X10	办公/零售业密度 (个/km ²)	14.90	15.94	0.26	87.86
X11	种族聚居情况 (0,1) ⁵	0.10	0.31	0.00	1.00
X12	流动人口比重 (%) ¹	6.81	7.55	0.00	65.59
X13	人均住房面积 (m ² /人)	8.89	1.71	7.53	15.10
X14	住房价格 (元/m ²)	6,686.54	3,361.22	1,400.00	18,000.00

注：¹ 中国的户籍制度将居民分为两类：常住人口和暂住人口。暂住人口是指来自农村地区的外来人口，在城市没有永久居住权，通常也称为“流动人口”。这里的人口密度为每平方公里的常住人口数。

² 就业率是全部适龄劳动力（男 18-60 岁，女 18-55 岁）中实际劳动的人口比重。

³ 抚养比是指非劳动力与实际劳动力之比。

⁴ 公共服务设施密度是指每平方公里拥有的政府机构、非盈利组织、教育设施、医疗设施、邮政通信设施数。

⁵ 种族聚居区是虚拟变量，用于确定某个乡镇、街道是否有少数民族（在北京主要为穆斯林）或外来人口集聚区。

表 7.3 主成分分析的特征值

主成分	特征值	方差比例	累计方差比例
1	<u>4.9231</u>	<i>0.3516</i>	0.3516
2	<u>2.1595</u>	<i>0.1542</i>	0.5059
3	<u>1.4799</u>	<i>0.1057</i>	0.6116
4	<u>1.2904</u>	<i>0.0922</i>	0.7038
5	0.8823	0.0630	0.7668
6	0.8286	0.0592	0.8260
7	0.6929	0.0495	0.8755
8	0.5903	0.0422	0.9176
9	0.3996	0.0285	0.9462
10	0.2742	0.0196	0.9658
11	0.1681	0.0120	0.9778
12	0.1472	0.0105	0.9883
13	0.1033	0.0074	0.9957
14	0.0608	0.0043	1.0000

表 7.4 社会区分析的因子载荷

变量	土地利用强度	邻里变量	社会经济地位	种族情况
公共服务设施密度	<u>0.8887</u>	0.0467	0.1808	0.0574
人口密度	<u>0.8624</u>	0.0269	0.3518	0.0855
就业率	<u>-0.8557</u>	0.2909	0.1711	0.1058
办公/零售业密度	0.8088	-0.0068	0.3987	0.2552
住房价格	<u>0.7433</u>	-0.0598	0.1786	-0.1815
抚养比	<u>0.7100</u>	0.1622	-0.4873	-0.2780
家庭规模	0.0410	<u>0.9008</u>	-0.0501	0.0931
流动人口比重	0.0447	<u>0.8879</u>	0.0238	-0.1441
人均住房面积	-0.5231	<u>0.6230</u>	-0.0529	0.0275
收入	0.1010	0.1400	<u>0.7109</u>	-0.1189
自然增长率	-0.2550	0.2566	<u>-0.6271</u>	0.1390
种族聚居情况	0.0030	-0.1039	-0.1263	<u>0.6324</u>
性别比	-0.2178	0.2316	-0.1592	<u>0.5959</u>
工厂密度	0.4379	-0.1433	0.3081	<u>0.5815</u>

表 7.5 社会区域特征（聚类结果）

类编号		乡镇街道数	平均因子得分			
5 种类型	9 种类型		土地利用强度	邻里变量	社会经济地位	种族情况
1	1. 郊区中等密度	21	-0.2060	0.6730	-0.6932	0.3583
2	2. 近郊区中等收入	23	-0.4921	-0.5159	-0.0522	0.4143
	4. 中心城中等收入	22	0.8787	-0.1912	0.5541	0.1722
	5. 远郊区中等收入	21	-0.8928	-0.8811	0.0449	-0.7247
3	3. 远郊区制造业和流动人口中心	6	-1.4866	2.0667	0.3611	0.1847
	9. 远郊区流动人口最密集地区	1	0.1041	5.7968	-0.2505	-1.8765
4	7. 中心城高收入	2	0.7168	0.9615	5.1510	-0.8112
	8. 中心城种族集聚区	1	1.8731	-0.0147	1.8304	4.3598
5	6. 中心城低收入	10	2.0570	0.0335	-1.1423	-0.7591

表 7.6 用虚拟变量描述环带和扇形区域

环带		扇形	
编号和位置	编码	编号和位置	编码
1. 第 2 环带以内	$x_2=x_3=x_4=0$	1. NE	$y_2=y_3=y_4=0$
2. 第 2、3 环带之间	$x_2=1, x_3=x_4=0$	2. SE	$y_2=1, y_3=y_4=0$
3. 第 3、4 环带之间	$x_3=1, x_2=x_4=0$	3. SW	$y_3=1, y_2=y_4=0$
4. 第 4 环带之外	$x_4=1, x_2=x_3=0$	4. NW	$y_4=1, y_2=y_3=0$

表 7.7 关于环带和扇形结构的回归分析 (n = 107)

因子		土地利用强度	邻里变量	社会经济地位	种族情况
环带模型	b_1	1.2980 ^{***} (12.07)	-0.1365 (-0.72)	0.4861 ^{**} (2.63)	-0.0992 (-0.51)
	b_2	-1.2145 ^{***} (-7.98)	0.0512 (0.19)	-0.4089 (-1.57)	0.1522 (0.56)
	b_3	-1.8009 ^{***} (-11.61)	-0.0223 (-0.08)	-0.8408 ^{**} (-3.16)	-0.0308 (-0.11)
	b_4	-2.1810 ^{***} (-14.47)	0.4923 (1.84)	-0.7125 ^{**} (-2.75)	0.2596 (0.96)
	R^2	0.697	0.046	0.105	0.014
扇形模型	c_1	0.1929 (1.14)	-0.3803 ^{**} (-2.88)	-0.3833 ^{**} (-2.70)	-0.2206 (-1.32)
	c_2	-0.1763 (-0.59)	-0.3511 (-1.52)	0.4990 [*] (2.01)	0.6029 [*] (2.06)
	c_3	-0.2553 (-0.86)	0.0212 (0.09)	1.6074 ^{***} (6.47)	0.4609 (1.58)
	c_4	-0.3499 (-1.49)	1.2184 ^{***} (6.65)	0.1369 (0.69)	0.1452 (0.63)
	R^2	0.022	0.406	0.313	0.051

* 显著性水平 0.05, ** 显著性水平 0.01, *** 显著性水平 0.001。