第五章 地理学中的经典统计分析方法

- 主成分分析
- 地统计分析方法
- 马尔可夫预测与趋势面分析

§5.1 主成分分析方法

在地理系统分析中，多变量问题是经常会遇到的，变量太多，无疑会增加分析问题的难度与复杂性。主成分分析就是寻找用较少的新变量代替原来较多的旧变量，而且使新变量尽可能多地保留原来较多信息的方法。

- 主成分分析的基本原理
- 主成分分析的计算步骤
- 主成分分析方法应用实例

★基本原理

- 假定有n个地理样本，每个样本共有p个变量，构成一个n×p阶的地理数据矩阵 （5.1.1）

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

当p较大时，在p维空间中考察问题比较麻烦。

- 如果记原来的变量指标为x_1，x_2，…，x_P，

它们的综合指标—新变量指标为z_1，z_2，…，z_m（$m \leqslant p$），则：

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \vdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{cases} \tag{5.1.2}$$

系数l_{ij}的确定原则：

① z_i与z_j（$i \neq j$；i，j=1，2，…，m）相互无关；

② z_1是x_1，x_2，…，xP的一切线性组合中方差最

大者，z_2是与z_1不相关的x_1，x_2，…，x_P的所有线性组合中方差最大者；……；z_m是与z_1，z_2，……，z_{m-1}都不相关的x_1，x_2，…，x_P的所有线性组合中方差最大者。

这样决定的新变量指标z_1，z_2，…，z_m分别称为原变量指标x_1，x_2，…，x_P的第一，第二，…，第m主成分。

主成分分析的主要任务就是确定每一个主成分z_i在原变量x_j上的载荷l_{ij}。

★计算步骤

◆ 计算相关系数矩阵：

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

式中r_{ij}（i，j=1，2，…，p）为原变量x_i与x_j的相关系数 。

◆ 计算特征值与特征向量：

① 解特征方程 $|\lambda I - R| = 0$，求出特征值，并使其按大小顺序排列，即 $\lambda_1 \geq \lambda_2 \geq \cdots; \geq \lambda_p \geq 0$ 。

② 分别求出对应于特征值 λ_i 的特征向量 $e_i(i=1,2,\cdots,p)$ 要求 $\|e_i\|=1$，即 $\sum_{j=1}^{p} e_{ij}^2 = 1$，其中 e_{ij} 表示向量的第j个分量。

◆计算主成分贡献率及累计贡献率：

①贡献率：

$$\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k}\quad(i=1,2,\cdots,p)$$

②累计贡献率：$\dfrac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}$ $(i=1,2,\cdots,p)$

◆计算主成分载荷

$$l_{ij}=p(z_i,x_j)=\sqrt{\lambda_i}e_{ij}(i,j=1,2,\cdots,p) \qquad (5.1.5)$$

◆ 各主成分的得分：

$$Z=\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix} \qquad (5.1.6)$$

★主成分分析方法应用实例

下面我们根据表3.1.1 给出的数据（见第3章），对耕地规模变化的驱动做主成分分析，其步骤如下：

（1）将表3.1.1中的数据做标准化处理，然后就爱那个它们代入公式（5.1.4），计算相关系数矩阵（见表5.1.1）

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	y
X_1	1									
X_2	−0.8283	1								
X_3	0.9678	−0.7033	1							
X_4	0.8902	−0.5202	0.9253	1						
X_5	0.8887	−0.5220	0.9247	0.9989	1					
X_6	0.8770	−0.5618	0.8817	0.9223	0.9126	1				
X_7	0.8306	−0.5823	0.8652	0.8734	0.8833	0.7670	1			
X_8	0.8495	−0.5109	0.8578	0.9075	0.8941	0.9333	0.6884	1		
X_9	−0.2179	0.3686	−0.1158	−0.0774	−0.0602	−0.243	0.102	−0.2876	1	
y	0.8187	−0.3599	0.8950	0.9468	0.9413	0.8815	0.7611	0.8985	−0.0393	1

（2）由相关系数矩阵计算特征值，一级各个主成分的贡献率和累计贡献率（见5.1.2）。由表（5.1.2）可知第一、第二主成分的累计贡献率已达到88%以上（大于85%），故只需要求出第一、第二主成分。

5.1.2 特征值与主成分贡献率

因子	特征值	贡献率%	累计贡献率%
1	6.7805	75.3391	75.3391
2	1.2098	13.4417	88.7809
3	0.6482	7.2022	95.9831
4	0.1779	1.9766	97.9597
5	0.0810	0.8995	98.8592
6	0.0508	0.5645	99.4237
7	0.0461	0.5126	99.9362
8	0.0052	0.0575	99.9937
9	0.0006	0.0063	100.0000

（3）对于特征值 λ_1 =6.7805，λ_2 =1.2098，λ_3 =0.6482分别求出特征向量 e_1, e_2, e_3，在用公式（5.1.5）计算各变量 $x_1, \ldots x_9$ 在主成分上的载荷矩阵（表5.1.3）

表5.1.3 主成分载荷

	主成分1	主成分2	主成分3
X_1	0.9715	-0.0894	-0.1842
X_2	-0.7047	0.3998	0.5735
X_3	0.9720	0.0531	-0.0913
X_4	0.9646	0.1650	0.1606
X_5	0.9621	0.1814	0.1411
X_6	0.9427	-0.0317	0.2212
X_7	0.8796	0.3002	-0.1823
X_8	0.9161	-0.0753	0.3168
X_9	-0.1937	0.9392	-0.2208

上述过程可以在SPSS\STATICTICA或MATLAB软件系统中实现。

分析：从表5.1.2可知，第一、二主成分的累计贡献率已达95.983%，完全符合主成分分析要求。由此进一步得出主成分载荷矩、阵主成分载荷矩阵是主成分与各个解释变量之间的相关系数。从表5.1.3中可以看，第一主成分与 x_1, x_3, x_4, x_5, x_6之间存在很大的正相关关系矩阵。第二主成分与具有大的正相关性。据此，且末绿洲耕地数量变化的驱动力可以归纳为经济发展、粮食生产驱动和社会系统推动这三驱动因素 。

分析部分

1.经济发展　从第一主成分的构成因子看，GDP、人均GDP和社会消费品零售总额占重要地位，经济系统中的动态因素对耕地数量变化的影响尤为显著。当前，且末县社会经济水平较低，经济基本处于传统的自然经济状态，农业人口占66%，农业总产值占GDP的86.62%（2004年）。这种情况下，人们开始寻找提高收入的新途径。因为可以获取直接物质产出及经济收入，耕地开发与投资成为人们最安全，理想的选择，直接驱动耕地数量变化。

2. 粮食生产驱动　第二主成分的主要构成因子是粮食总产量（x_9），与第二主成分的相关系数达0.9392。粮食生产驱动有两重意义，一方面，粮食是人类生存的最基本的要素，耕地是粮食生产的最终源泉。且末绿洲在人口迅速增长的背景下，保证最基本的粮食供给，维持集体和个人安全的前提是保证一定数量的耕地；另一方面，随着农业生产中科技因素的强化，土地生产率的提高，食物产品价值增长，食物生产变成了且末农村经济的一个重要组成部分。

§5.2 趋势面分析方法

趋势面分析，是利用数学曲面模拟地理系统要素在空间上的分布及变化趋势的一种数学方法。

一，基本原理

趋势面分析运用最小二乘法拟合一个二维非线性函数，模拟地理要素在空间上的分布规律，展示地理要素在地域空间上的变化趋势。

设某地理要素的实际观测数据为 $z_i(x_i, y_i)$ $(i = 1, 2, \ldots, n)$ 趋势面拟合值为 $\hat{z}_i(x_i, y_i)$ ，则有

$$z_i(x_i, y_i) = \hat{z}_i(x_i, y_i) + \varepsilon_i \qquad (5.2.1)$$

式中：ε_i 即为剩余值（残差值）。

趋势面分析的核心就是从实际观测值出发推算趋势面，一般采用回归分析方法，使得残差平方和趋于最小，即

$$Q = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} [z_i(x_i, y_i) - \hat{z}_i(x_i, y_i)]^2 \to \min$$

多项式趋势面的形式

一次趋势面模型

$$z = a_0 + a_1 x + a_2 y \qquad (5.2.2)$$

二次趋势面模型

$$z = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 xy + a_5 y^2 \qquad (5.2.3)$$

三次趋势面模型

$$z = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 xy + a_5 y^2 + a_6 x^3 + a_7 x^2 y + a_8 xy^2 + a_9 y^3 \qquad (5.2.4)$$

趋势面模型参数的估计

趋势面模型参数的估计实际上就是根据观测值 z_i, x_i, y_i $(i=1,2,...,n)$确定多项式的系数 $a_0, a_1, ..., a_p$

使残差平方和最小，其具体过程如下：

将多项式回归（非线性模型）模型转化为多元线性回归模型。

若令

$$x_1 = x, x_2 = y, x_3 = x^2, x_4 = xy, x_5 = y^2, \cdots$$

则

这样就 $\hat{z} = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p$ 多项式回归（非线性模型）模型转化为多元线性回归模型。其残差平方和为

$$Q = \sum_{i=1}^{n} [z_i - \hat{z}_i]^2 = \sum_{i=1}^{n} [z_i - (a_0 + a_1 x_{1i} + a_2 x_{2i} + \cdots + a_p x_{pi})]^2$$

根据最小二乘法的原理 ，求 Q 对 $a_0, a_1, ..., a_p$ 的偏导数，并令其等于0，择地正规方程组

$$\begin{cases} na_0 + a_1\sum_{i=1}^{n} x_{1i} + \cdots + a_p\sum_{i=1}^{n} x_{pi} = \sum_{i=1}^{n} z_i \\ a_0\sum_{i=1}^{n} x_{1i} + a_1\sum_{i=1}^{n} x_{1i}x_{1i} + \cdots + a_p\sum_{i=1}^{n} x_{pi}x_{1i} = \sum_{i=1}^{n} x_{1i}z_i \\ \cdots\cdots\cdots \\ a_0\sum_{i=1}^{n} x_{pi} + a_1\sum_{i=1}^{n} x_{1i}x_{pi} + \cdots + a_p\sum_{i=1}^{n} x_{pi}x_{pi} = \sum_{i=1}^{n} x_{pi}z_i \end{cases}$$

则以上方程组变为

$$X^{T}XA = X^{T}Z \qquad (5.2.7)$$

二，趋势面模型的适度检验

（一）趋势面拟合适度的R^2检验

一般用变量z的总离差平方和中回归平方和所占的比重表示回归模型的拟合优度总离差平方和等于回归平方和与剩余平方和之和。即：

$$SS_T = \sum_{i=1}^{n}(z_i - \hat{z}_i)^2 + \sum_{i=1}^{n}(\hat{z}_i - \bar{z})^2 = SS_D + SS_R$$

SS_R越大（或 SS_D 越小）就表示因变量与自变量的关系越密切，回归的规律性越强、效果越好。记：

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_D}{SS_T} \qquad (5.2.9)$$

R^2越大，趋势面的拟合度就越高

（二）趋势面拟合适度的显著性*F*检验

趋势面适度的F检验，是对趋势面回归模型整体的显著性检验。即：

$$F = \frac{SS_R / p}{SS_D / n - p - 1} \qquad (5.2.10)$$

检验的在显著性水平 α 下，查 F 分布表得 F_α

若计算的F值大于临界值 F_α 则认为趋势面方程显著；反之则不显著。

（三）趋势面适度的逐次检验

趋势面适度的逐次检验的基本思想与方法可以如下概括：

(1)求出较高次多项式方程的回归平方和与较低次多项式方程的回归平方和之差；

(2)将此差除以回归平方和的自由度之差，得出由于多项式次数增高所产生的回归均方差；

(3)将此均方差除以较高次多项式的剩余均方差，得出相继两个阶次趋势面模型的适度性比较检验值F。

三，趋势面分析应用实例

某流域1月份降水量与各观测点的坐标位置数据如表5.2.2所示。下面，我们以降水量为因变量z，地理位置的横坐标和纵坐标分别为自变量x、y，进行趋势面分析，并对趋势面方程进行适度F检验。

表5.2.2　流域降水量及观测点的地理位置数据

序号	降水量Z/mm	横坐标 $x/10^4$ m	纵坐标 $y/10^4$ m
1	27.6	0	1
2	38.4	1.1	0.6
3	24	1.8	0
4	24.7	2.95	0
5	32	3.4	0.2
6	55.5	1.8	1.7
7	40.4	0.7	1.3
8	37.5	0.2	2
9	31	0.85	3.35
10	31.7	1.65	3.15
11	53	2.65	3.1
12	44.9	3.65	2.55

解题步骤

- 建立趋势面模型

（1）首先采用二次多项式进行趋势面拟合，用最小二乘法求得拟合方程为

$$z = 5.998 + 17.438x + 29.787y - 3.558x^2 + 0.357xy - 8.070y^2$$

$$R^2 = 0.839, F = 6.236$$

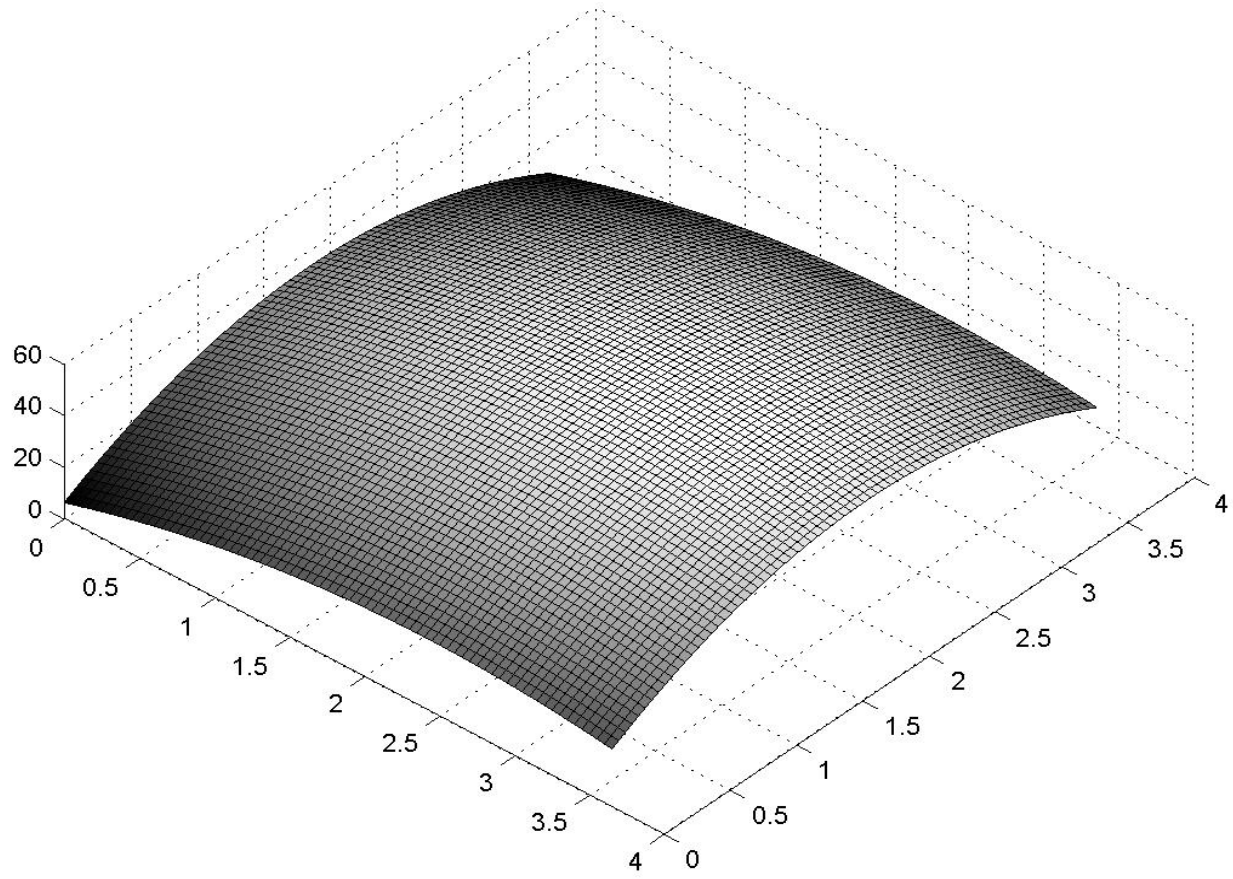

图5.2.1　某流域降水量的二次多项式趋势面

(2)再采用三次趋势面进行拟合，用最小二乘法求得拟合方程为

$$z = -48.810 + 37.557x + 130.130y + 8.389x^2 - 33.166xy$$

$$-62.740y^2 - 4.133x^3 + 6.138x^2y + 2.566xy^2 + 9.785y^3$$

$$R^2 = 0.965, F = 6.054$$

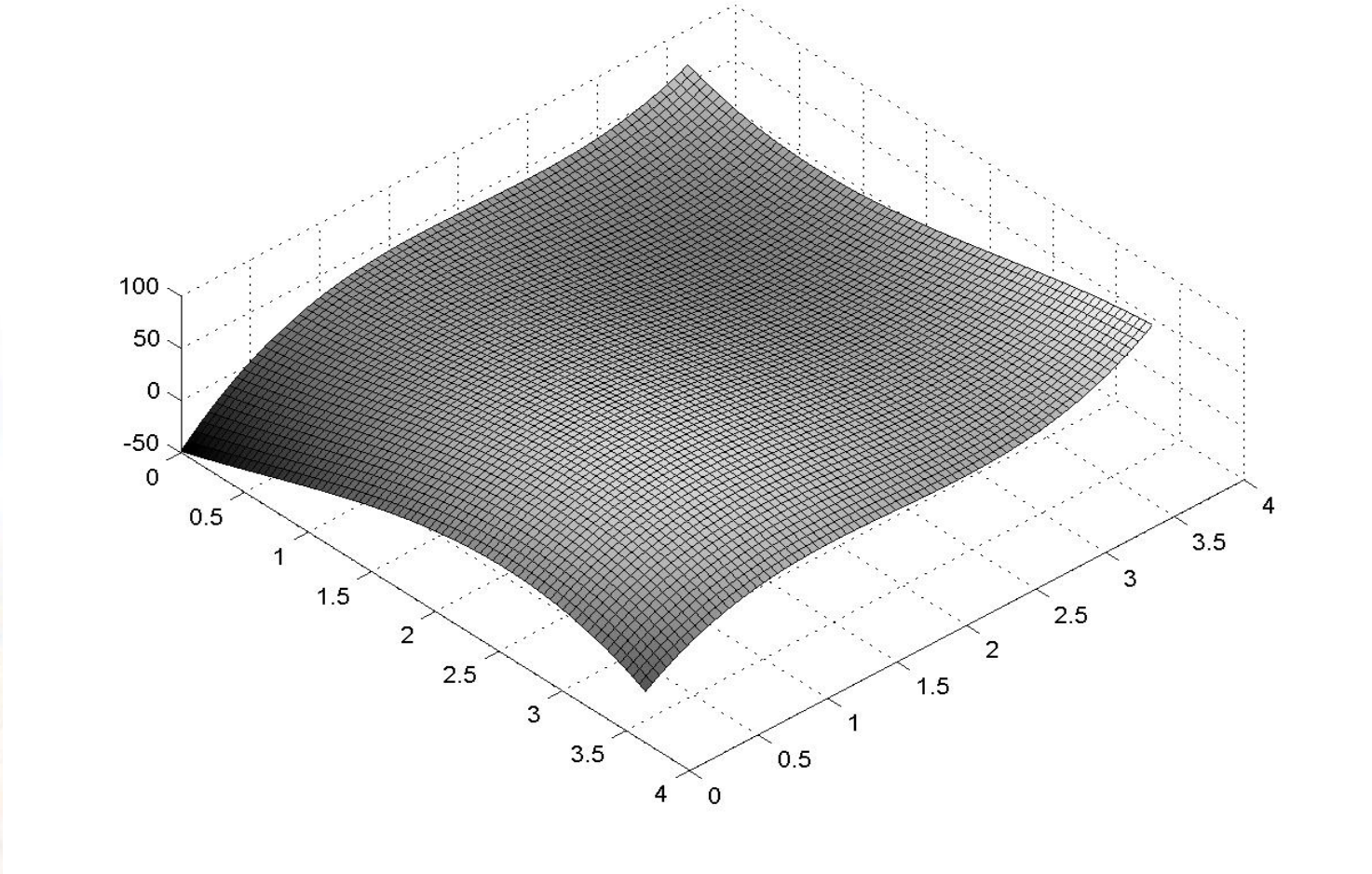

图5.2.2　某流域降水量的三次多项式趋势面

§5.3 马尔可夫预测方法

马尔可夫预测方法---对于地理事件的预测，不仅要能够指出事件发生的各种可能结果，而且还必须给出每一种结果出现的概率，说明被预测的事件在预测期内出现每一种结果的可能性程度。

一，几个基本概念

（一）状态、状态转移过程与马尔可夫过程

（1）状态：指某一事件在某个时刻（或时期）出现的某种结果。

（2）状态转移过程。事件的发展，从一种状态转变为另一种状态，称为状态转移。

（3）马尔可夫过程。在事件的发展过程中，若每次状态的转移都仅与前一时刻的状态有关，而与过去的状态无关，则这样的状态转移过程就称为马尔可夫过程。

（二）状态转移概率与状态转移概率矩阵

（1）状态转移概率。在事件的发展变化过程中，从某一种状态出发，下一时刻转移到其它状态的可能性，称为状态转移概率。由状态E_i转为状态E_j的状态转移概率 $P(E_i \to E_j)$ 是

$$P(E_i \to E_j) = P(E_j / E_i) = P_{ij} \qquad (5.3.1)$$

（2）状态转移概率矩阵：假定某一个事件的发展过程有n个可能的状态，即E_1，E_2，…，E_n。记为从状态Ei转变为状态Ej的状态转移概率 $P(E_i \to E_j)$，则矩阵

$$P=\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{bmatrix}$$

称为状态转移概率矩阵。

二，马尔可夫预测法

为了运用马尔可夫预测法对事件发展过程状态出现的概率进行预测，还需要在介绍一个名词，即专题概率$\pi_j(k)$。

状态 E_j 的概率显然：

$$\sum_{j=1}^{n}\pi_j(k)=1 \qquad (5.3.6)$$

计算公式 为：

$$\pi_j(k)=\sum_{i=1}^{n}\pi_j(k-1)P_{ij} \qquad (j=1,2,\cdots,n) \tag{5.3.7}$$

逐次计算状态概率的递推公式为：

$$\begin{cases} \pi(1)=\pi(0)P \\ \pi(2)=\pi(1)P=\pi(0)P^1 \\ \quad\vdots \\ \pi(k)=\pi(k-1)P=\cdots=\pi(0)P^k \end{cases} \tag{5.3.8}$$