一.相关分析方法

1.相关系数：揭示地理要素（变量）之间的相互关系密切程度。

(1)相关系数的计算和检验

①对于两个变量 x 和 y 相关系数被定义为

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

x_i,y_i 为样本值。

如果记:

$$L_{xy}=\sum_{i=1}^{n}(xi-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)$$

$$L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{1}{x}\left(\sum_{i=1}^{n}x_i\right)^2$$

$$L_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2$$

可将公式简化为

$$r_{xy}=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$$

r_{xy} 取值介于[-1,1]之间，其绝对值越接近 1，表示两变量关系越密切。

②相关系数矩阵

若有 $x_1,x_2\cdots x_n$ 个变量, 对其中任意两个变量计算相关系数, 得到多变量的相关系数矩阵。

$$R=\begin{bmatrix}r_{11} & r_{12}\cdots r_{1n}\\ r_{21} & r_{22}\cdots r_{2n}\\ M & M & M\\ r_{n1} & r_{n2}\cdots r_{nn}\end{bmatrix}$$

R 为对称矩阵，$r_{ii}=1,r_{ij}=r_{ji}$

(2)相关系数的检验 通过检验才能知道可信度

相关系数的检验是在给定的置信水平下，通过查相关系数检验的临界值表来完成。

如下表:

检验相关系数ρ=0 的临界值（r_a）表

$$p\{|r|>r_a\}=\alpha$$

f \ α	0.10	0.05	0.02	0.01	0.0001
1	0.98769	0.99692	0.999507	0.999877	0.9999988
2	0.90000	0.95000	0.9800	0.99000	0.999000
3	0.8054	0.8783	0.93433	0.95873	0.991160
4	0.7293	0.8114	0.8822	0.91720	0.97406
5	0.6694	0.7545	0.8329	0.8745	0.95047
6	0.6215	0.7067	0.7887	0.8343	0.92493
7	0.5822	0.6664	0.7493	0.7977	0.8982
8	0.5494	0.6319	0.7155	0.7646	0.8721
9	0.5214	0.6021	0.6851	0.7348	0.8471
10	0.4973	0.5760	0.6581	0.7079	0.8233

F 称为自由度，其值为 n-2，α 代表不同的置信水平。

公式 p= { | r | > ra} =a 的意思是当所计算的相 关系数 r 的绝对值大于在 a 水平下的临界值 ra 时，两要素不相关（即 ρ=0) 的可能性只有 a

一般而言，当 | r | < $r_{0.1}$ 时，则认为两要素不相关，这时的样本相关系 数就不能反映两要素之间的关系。

2. 秩相关系数的计算与检验

（1）计算

秩相关系数也是描述两变量之间相关程度的一种统计指标，不过在计算方法上，与前述相关系数 的计算有所不同。秩相关系数是将两变量的样本值按数值的大小顺序排列 位次，以各要素样本值的位次代替实际数据而求得的一种统计量。

设两个变量 x 和 y 有 n 对样本值，令 R1 代表变量 x 的序号（或位次)， R_2 代表变量 y 的序号（或位次)，定义 $d_i^2 = (R_{1i} - R_{2i})^2$ 代表变量 x 和 y 的同一 组样本位次差的平方，那么要素 x 与 y 之间的秩相关系数（r'_{xy})被定义

$$r'_{xy} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

举个例子

省(市、区)	总人口(x)及其位次		社会总产量(y)及其位次		位次差的平方 $d_i^2 = (R_{1i} - R_{2i})^2$
	人口数(万人)	位次 R_1	总产值(亿元)	位次 R_2	
北京	960	25	482.10	14	121
天津	808	26	431.92	15	121
河北	5548	7	715.30	10	9
山西	2627	19	404.40	17	4
内蒙	2007	22	252.53	23	1
辽宁	3686	12	1076.53	4	64
吉林	2298	20	422.24	16	16

（2）秩相关系数的检验—同样通过查表

n	显著水平 a	
	0.05	0.01
4	1.000	
5	0.900	1.000
6	0.829	0.943
7	0.714	0.893
8	0.643	0.833
9	0.600	0.783
10	0.564	0.746
12	0.506	0.712
14	0.456	0.645

3.多变量间的相关程度的检验

(1)偏相关系数的计算和检验

在多个变量的系统当中，当研究一个变量对另一个变量的影响时，把其他变量的影响视为常数，而单独研究两个变量之间的相互关系的密切程度时，称为偏相关。

①计算

偏相关系数可以用单向关系数来计算

假设有 3 个变量 x_1,x_2,x_3,其计算公式如下:

$$r_{12\cdot 3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(r-r_{23}^2)}}$$

$$r_{13\cdot 2}=\frac{r_{13}-r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$

$$r_{23\cdot 1}=\frac{r_{23}-r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

$r_{12\cdot 3}$ 代表变量 x_3 保持不变的情况下，x1 与 x2 的偏相关程度。

同样有四个或更多的偏相关系数计算公式。

偏相关系数的性质:

①分布范围在-1 到 1 之间

②绝对值越大，偏相关程度越大

③绝对值必小于等于同一系列所得的复相关系数

(2)偏相关系数的检验

采用 t 检验法

$$t=\frac{r_{12\cdot 34\cdots m}}{\sqrt{1-r_{12\cdot 34\cdots m}^2}}\sqrt{n-m-1}$$

n 为样本数，m 为 自变量个数（通过后面的复相关系数应为变量数减 1）

通过查 t 分布表 进行检验

（3）复相关系数的计算和检验

一个变量的变化往往受多种变量的综合作用和影响，而单相关或偏相关分析 的方法都不能反映各变量的综合影响。几个变量与某一个变量之间的复相关程度，可用复相关系数来测定。

①计算

设 Y 为因变量，X_1，X_2，…，X_k 为自变量，则将 Y 与 X_1，X_2，…，X_k 之 间的复相关系数记为 $R_{y\cdot 12\cdots k}$。其计算公式如下

当有两个自变量时

$$R_{y\cdot 12}=\sqrt{1-(1-r_{y1}^2)(1-r_{y2\cdot 1}^2)}$$

r_{y1} 为单向关系数，$r_{y2.1}$ 为偏相关系数

一般的，当有 k 个相关系数时

$$R_{y\cdot 12-k}=\sqrt{1-(1-r_{y1}^2)(1-r_{y2\cdot 1}^2)\cdots[1-r_{yk\cdot 12\cdots(k-1)}^2]}$$

复相关系数的性质:

a.复相关系数介于 0 到 1 之间

b.复相关系数越大，表明变量之间的密切程度越密切

c.复相关系数必大于等于单向关系数的绝对值

②检验

采用 F 检验法

$$F=\frac{R_{y\cdot 12\cdots k}^2}{1-R_{y\cdot 12\cdots k}^2}\times\frac{n-k-}{k}$$

应为 n-k-1

查 F 检验的临界值表进行检验

二.回归分析方法

用一定的函数形式表达某一变量和其他变量之间的相互关系

1.一元线性回归模型

$$\hat{y} = \hat{a} + \hat{b}X$$

1)参数 a 和 b 的最小二乘估计

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$b = \frac{L_{xy}}{Lx_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

建立一元线性模型回归模型的过程就是用变量 x 和 y 的实际观测值确定参数 a 和 b 的过程

2）一元线性回归模型的显著性检验

在回归分析中，y 的 n 次观测值 y1，y2…，yn 之间的差异，可用观测 值 与其平均值 的离差平方和来表示，它被称为总的离差平方和，记为

$$S_{总} = L_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$S_{总} = L_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$= Q + U$$

其中 $Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 称为误差平方和，$U = \sum_{i=1}^{N}(\hat{y}_i - \bar{y}i)^2$ 称为回归平方和。误差平方和应尽可能小，回归平方和应尽可能大。

用统计量 $F = U / \frac{Q}{n-2}$ 衡量回归模型的效果。F 服从自由度 f1=1 和 f2=n-2 的 F 分布。

即 F~F（1，n-2） 查 F 分布表进行检验

2.多元线性回归模型

回归方程为

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

$b_1, b_2, b_3 \cdots b_k$ 称为偏回归系数

如果引入以下矩阵

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ 1 & x_{13} & x_{23} & \cdots & x_{k3} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} \quad b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix}$$

$$b = A^{-1}B = (X^T X)^{-1} X^T Y$$

2）多元线性回归模型显著性 检验

在多元线性回归分析中，各平方和的自由度略有不同，回归平方和 U 的自由度等于自变量的个数 K，而剩余平方和的自由度等于 n-K-1，所以 F 统计量为

$$F=\frac{U/K}{Q/(n-K-1)}$$

算出 F 后查 F 分布表进行检验

3.非线性回归模型的建立方法

1）将非线性关系线性化

譬如：

(1)对于指数曲线 $y=de^{bx}$，令 $y'=\ln y$，$x'=x$，就可以将其转化为直线形式：$y'=a+bx'$，其中，$a=\ln d$；

(2)对于对数曲线 $y=a+b\ln x$，令 $y'=y$，$x'=\ln x$，就可以将其转化为直线形式：$y'=a+bx'$；

(3)对于幂函数曲线 $y=dx^{b}$，令 $y'=\ln y$，$x'=x$，就可以将其转化为直线形式：$y'=a+bx'$，其中，$a=\ln d$；

(4) 对于双曲线 $\frac{1}{y}=a+\frac{b}{x}$，令 $y'=\frac{1}{y}$，$x'=\frac{1}{x}$，就可以将其转化为直线形式：$y'=a+bx'$；

(5) 对于S型曲线 $y=\frac{1}{a+be^{-x}}$，令 $y'=\frac{1}{y}$，$x'=e^{-x}$，就可以将其转化为直线形式：$y'=a+bx'$；

(6)对于幂函数乘积：

$$y=dx_1^{\beta 1}-x_2^{\beta 2}\cdots,\ x_k^{\beta k}$$

只要令 $y'=\ln y$，$x_1'=\ln x1$，$x_2'=\ln x_2$，$\cdots$，$x_k'=\ln x_k$，就可以将其转化为直线形式：

$$y'=\beta_0+\beta_1x_1'+\beta_2x'+\cdots+\beta_kx_k'$$

上式中，$\beta_0=\ln d$；

(7)对于对数函数和：$y=\beta_0+\beta_1\ln x_1+\beta_2\ln x_2+\cdots+\beta_k\ln x_k$

只要令 $y'=y$，$x_1'=\ln x_1$，$x_2'=\ln x_2$，$\cdots$，$x_k'=\ln x_k$，就可以将其化为线性形式：

$$y'=\beta_0+\beta_1x_1'+\beta_2x_2'+\cdots+\beta_kx_k'$$

到建立非线性回归模型的一般方法：首先通 过适当的变量替换将非线性关系线性化，然后再用线性回归分析方法建立新 变量下的线性回归模型，通过新变量之间的线性相关关系反映原来变量之间 的非线性相关关系。

三.时间序列分析

时间序列：是变量的数据按照时间顺序变动排列形成的一种数列，反映了变量随时间变化的发展过程

1.平滑预测法　为消除偶然因素对变量的影响，对数据的一种处理

平滑预测法又主要分为三类

1）移动平均法

设某一时间序列 $y_1,y_2,\cdots y_t$,则 y_{t+1} 的预测值为前 n 个值的平均值。

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

2）滑动平均法

Y_t 的预测值为左侧 l 个值与右侧 l 个值，以及 y_t 自身的平均值　l 称为单侧平滑时距

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

举个例子　　下面应为三点移动

表 3.3.1　中国 1990—1999 年农业总产值及平滑结果

序号	年份	农业总产值	移动平均法		滑动平均法	
			三点移动	五点移动	三点滑动	五点滑动
1	1990	7 662.1				
2	1991	8 157.0			8 301.26	
3	1992	9 084.7			9 412.47	10 329.96
4	1993	10 995.5	8 301.26		11 943.57	12 865.72
5	1994	15 750.5	9 412.47		15 695.63	15 705.06
6	1995	20 340.9	11 943.57	10 329.96	19 481.70	18 645.8
7	1996	22 353.7	15 695.63	12 865.72	22 161.00	21 355.08
8	1997	23 788.4	19 481.70	15 705.06	23 561.33	23 108.8
9	1998	24 541.9	22 161.00	18 645.8	24 283.13	
10	1999	24 519.1	23 561.33	21 355.08		
11	2000		24 283.13	23 108.8		

3）指数平滑法

（1）一次指数平滑法

按照距离预测期的远近给与大小不同的权数

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

α称为平滑系数　具体应用时通过经验和 试算取值

举个例子：

表 3.3.2　某城市近 6 年用水量数据

年份	1994	1995	1996	1997	1998	1999
产量/10^6t	211.30	260.18	209.10	248.79	241.00	250.00

解：取 $\alpha=0.5$，将表 3.3.2 中的数据代入公式(3.3.3)计算

$$\begin{aligned}\hat{y}_{2000} &= \alpha y_{1999}+\alpha(1-\alpha)y_{1998}+\alpha(1-\alpha)^2 y_{1997}+\alpha(1-\alpha)^3 y_{1996}+\\ &\quad \alpha(1-\alpha)^4 y_{1995}+\alpha(1-\alpha)^5 y_{1994}\\ &=0.5\times 250.00+0.5\times 0.5\times 241.00+0.5\times 0.5^2\times 248.79+\\ &\quad 0.5\times 0.5^3\times 209.10+0.5\times 0.5^4\times 260.18+0.5\times 0.5^5\times 211.30\\ &=240.85(10^6\ \text{t})\end{aligned}$$

（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

二次指数平滑 法的预测公式为

$$\hat{y}_{t+k}=a_t+b_tT$$

T 为从基数 t 到预测期的期数

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

$$b_t=\frac{\alpha}{1-\alpha}(S_t^{(1)}-S_t^{(2)})$$

$S_t^{(1)}$代表一次指数平滑值　$S_t^{(2)}$代表二次指数平滑值　即对 一次指数平滑再做指数平滑

2.趋势线预测法

只简单说了三种最常用的趋势线

①直线型趋势线

$y_t=a+bt$

②指数型趋势线

$y_t=ab^t$

③抛物线型趋势线

$y_t=a+bt+ct^2$

系数的求解用回归分析中的方法

3.季节性预测法

书中讲，经过长期研究，可以讲时间序列分解为长期趋势 T，季节变动 S，循环变动 C 和不规则变动 I 四种

季节变动预测法的基本原理是用经过季节系数的趋势线进行预测。基本步骤为:

①将原序列求移动平均

②将原序列 y 除以对应的趋势方程值

③得到校正系数

④进行预测

举个例子:

表 3.3.3　某市 1995—1999 年各季度客流量　　单位：10^4 人次

季　度	1	2	3	4
1995	1 317.9	1 372.4	1 275.0	1 403.9
1996	1 434.8	1 444.1	1 296.1	1 462.6
1997	1 343.8	1 434.7	1 284.1	1 454.1
1998	1 262.8	1 477.9	1339.7	1 424.3
1999	1 338.7	1 373.0	1 248.4	1 338.4
均值	1 339.6	1 420.4	1 288.7	1 416.7

若要预测 2000 各季的客流量

①先对序列求三次滑动平均值

表 3.3.4　三次滑动平均值

季　度	1	2	3	4
1995	—	1 321.8	1 350.4	1 371.2
1996	1 427.6	1 391.7	1 400.9	1 367.5
1997	1 413.7	1 354.2	1 391.0	1 333.7
1998	1 398.3	1 360.1	1 414.0	1 367.6
1999	1 378.7	1 320.0	1 319.9	—

②用三次指数平滑法预测模型系数　α选用 0.3（经多次尝试后选用）　下表为所求三次指数平滑系数表

表 3.3.5　三次指数平滑模型系数

季　度		y_t	S_1	S_2	S_3	a_t	b_t	c_t
1995	1	1 317.90	1 317.90	1 317.90	1 317.90	1 317.90	.00	.00
	2	1 372.40	1 334.25	1 322.81	1 319.37	1 353.69	37.32	3.47
	3	1 275.00	1 316.48	1 320.91	1 319.83	1 306.54	-11.12	-1.12
	4	1 403.90	1 342.70	1 327.45	1 322.12	1 367.87	50.77	4.69
1996	1	1 434.80	1 370.33	1 340.31	1 327.58	1 417.64	103.00	9.44
	2	1 444.10	1 392.46	1 355.96	1 336.09	1 445.59	131.29	11.88
	3	1 296.10	1 363.55	1 358.23	1 342.73	1 358.69	36.49	2.89
	4	1 462.60	1 393.27	1 368.74	1 350.54	1 424.13	94.90	8.43
1997	1	1 343.80	1 378.43	1 371.65	1 356.87	1 377.21	39.66	3.23
	2	1 434.70	1 395.31	1 378.75	1 363.43	1 413.11	68.25	5.97
	3	1 284.10	1 361.95	1 373.71	1 366.52	1 331.24	-23.57	-2.58
	4	1 454.10	1 389.59	1 378.47	1 370.10	1 403.46	43.18	3.83
1998	1	1 262.80	1 351.55	1 370.40	1 370.19	1 313.64	-53.37	-5.17
	2	1 477.90	1 389.46	1 376.12	1 371.97	1 411.99	43.70	4.06

66　　第 3 章　统计分析方法

续表

季　度		y_t	S_1	S_2	S_3	a_t	b_t	c_t
	3	1 339.70	1 374.53	1 375.64	1 373.07	1 369.74	.38	-.07
	4	1 424.30	1 389.46	1 379.79	1 375.09	1 404.10	34.01	3.10
1999	1	1 338.70	1 374.23	1 378.12	1 376.00	1 364.33	-8.15	-.88
	2	1 373.00	1 373.86	1 376.84	1 376.25	1 367.31	-7.67	-.77
	3	1 248.40	1 336.22	1 364.65	1 372.77	1 287.48	-92.13	-8.58
	4	1 338.40	1 336.88	1 356.32	1 367.84	1 309.51	-71.20	-6.41

求得 1999 年 4 季度的 a_t,b_t,c_t 系数

③求预测模型

$$\hat{y}_{t+T} = 1\,309.51 - 71.20T - 6.41T^2$$

④求季节性指标　用表 3.3.3 的数据除以表 3.3.4 中的各元素　得到季节系数　再对各季季节性系数平均得到季节性指标

表 3.3.6　季节系数与季节性指标

季　度	1	2	3	4
1995	/	1.038 3	0.944 2	1.023 8
1996	1.005 0	1.037 7	0.925 2	1.069 5
1997	0.950 6	1.059 4	0.923 1	1.090 3
1998	0.903 1	1.086 6	0.947 5	1.041 5
1999	0.971 0	1.040 2	0.945 8	/
季节性指标	0.957 42	1.052 4	0.937 2	1.056 3

季节性指标和理论上应为 4　（一共四个季度）
现在为 4.0033　对其调整　求季节系数θ=4/4.0033=0.9992
让季节性指标乘以θ
⑤求预测值　用③中的模型

$\hat{y}_{20+1} = 1\,309.51 - 71.20\times 1 - 6.41\times 1^2 = 1\,231.90\times 10^4$ 人次
$\hat{y}_{20+2} = 1\,309.30 - 71.20\times 2 - 6.41\times 2^2 = 1\,141.47\times 10^4$ 人次
$\hat{y}_{20+3} = 1\,309.30 - 71.20\times 3 - 6.41\times 3^2 = 1\,038.22\times 10^4$ 人次
$\hat{y}_{20+4} = 1\,309.30 - 71.20\times 4 - 6.41\times 4^2 = 922.15\times 10^4$ 人次

将上述预测值乘以调整后的季节性指标　得到最终的预测值
4.自回归模型
当一个变量按时间顺序的观察值之间具有依赖关系或自相关性时，就可以建立该变量的自回归模型
1）时间序列的自相关性判断
设 $y_1 \cdots y_n$ 共有 n 个观察值。把前后相邻两期的观察值一一成对，便有 n-1 对数据，即(y1,y2)，…(y_{n-1},y_n) 其一阶自相关系数 r_1 为

$$r_1 = \frac{\sum_{t=1}^{n-1}(y_t - \bar{y}_t)(y_{t+1} - \bar{y}_{t+1})}{\sqrt{\sum_{t=1}^{n-1}(y_t - \bar{y}_t)^2 \cdot \sum_{t=1}^{n-1}(y_{t+1} - \bar{y}_{t+1})^2}}$$

一般的，k 阶自相关系数 r_k 为

$$r_k = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y}_t)(y_{t+k} - \bar{y}_{t+k})}{\sqrt{\sum_{t=1}^{n-k}(y_t - \bar{y}_k)^2 \cdot \sum_{t=1}^{n-k}(y_{t+k} - \bar{y}_{t+k})^2}}$$

2）自相关模型的建立
一阶线性自回归预测模型为

$$y_t = \psi_0 + \psi_1 y_{t-1} + \xi$$

一般的 p 阶线性自回归模型为

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t$$

参数值通过最小二乘法估计获得

四.系统聚类分析方法

1.数据处理 目的：消除不同的单位和量纲

①总和标准化

$$x'_{ij} = x_{ij} / \sum_{i=1}^{m} x_{ij} \qquad \begin{pmatrix} i = 1,\ 2,\ \cdots,\ m \\ j = 1,\ 2,\ \cdots,\ n \end{pmatrix}$$

m 为样本数　　x_i 表示第 i 个变量 x_{ij} 表示 x_i 变量的第 j 个取值　将 xi 变量的样本值除以 xi 的总样本和

②标准差标准化

$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad \begin{pmatrix} i = 1,\ 2,\ \cdots, m \\ j = 1, 2,\ \cdots,\ n \end{pmatrix}$ S_j 为变量 x_i 的 标准差

③极大值标准化

$$x'_{ij} = \frac{x_{ij}}{\max\{x_{ij}\}} \qquad \begin{pmatrix} i = 1,\ 2,\ \cdots,\ m \\ j = 1,\ 2,\ \cdots,\ n \end{pmatrix}$$

④极差的标准化

$$x_{ij} = \frac{x_{ij} - \min\{x_{ij}\}}{\max\limits_{i}\{x_{ij}\} - \min\{x_{ij}\}} \qquad \begin{pmatrix} i = 1,\ 2,\ \cdots,\ m \\ j = 1, 2,\ \cdots,\ n \end{pmatrix}$$

2.距离的计算

距离是系统聚类分析的依据和基础

常见的距离有：

①绝对值距离

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

②欧氏距离

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

③明科夫斯基距离

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

在进行聚类前，先采用一种距离方法计算距离矩阵 如下

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

表示每一个数表示第 i 个变量与第 j 个变量之间的距离

3.直接聚类法

是根据距离矩阵的结构一次并类得到结果，是一种简便的聚类方法。

它先把各个分类对象单独视为一类，然后根据距离最小的原则，依次选出一对分类对象，并成新类

如果一个分类对象已归于一类 ，则把另一个也归于该类；如果一对分类对象正好属 于已归的两类，则把这两类并为一类。

每一次归并，都划去该对象所在的列 与列序相同的行。那么，经过 m-1 次就可以把全部分类对象归为一类，这样 就可以根据归并的先后顺序作出聚类分析的谱系图。

直接 聚类方法简单但是直接划去行和列难免有信息损失

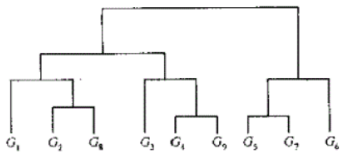


图 3.4.1 直接聚类谱系图

4.最短距离聚类法

最短距离法，是在原来的 m×m 距离矩阵的非对角元素中找出 $d_{pq}=\min\{d_{ij}\}$，把分类对象 Gp 和 Gq 归并为一新类 Gr，然后按计算公式： $d_{rk}=\min\{d_{pk}, d_{qk}\}$ (k≠p，q) 计算原来各类与新类之间的距离，这样就得到一个新的（m-1)阶的距离 矩阵；再从新的距离矩阵中选出最小的 d_{ij}，把 Gi 和 Gj 归并成新类；再计算 各类与新类的距离，这样一直下去，直至各分类对象被归为一类为止。

比如以一共有 x1 到 x9 九个变量 距离矩阵为 9*9 的矩阵 矩阵中最小值为 x_{34} 那么将变量 3 4 归为一类，为第 10 类，其余七类与第十类的距离按照上述 $\min\{d_{pk}, d_{qk}\}$ 计算，得到一个 8*8 视为距离矩阵 重复直至只有一类

5.最远距离聚类法

与最短距离聚类相似，但是选用 $\max\{d_{pk}, d_{qk}\}$

五主成分分析方法

1.基本原理

主成分分析是把原来多个变量化为少数几个综合指标的一种统计分析方 法，是一种降维处理技术。

用较少的综合指标代替原来较多的变量指标

综合指标（新变量）的选取：取原来变量的线性组合，适当调整组合系数，使新的变量之间相互独立且代表性最好。

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

p 个变量 n 个样本

系数 l_{ij} 由以下原则来决定

①z_i 与 z_j 相互无关

②z1 是 x1,x2…xp 的一切线性组合中方差最大者，z2 是与 z1 不相关的 x1,x2…xp 的所有线性组合中方差最大者；…

新变量 z1，z2，zm 分别称为 原变量指标的第一，第二，……第 m 主成分

2.计算步骤

①计算相关系数矩阵 R

②计算特征值与特征向量

首先解特征方程|λI-R|=0 I 为单位阵 求出λ_i 并按大小排序为$\lambda_1, \lambda_2 \cdots\cdots \lambda_p$ 求出对应的特征向量 e_i 并且$||e_i||=1$

③计算主成分贡献率及累计贡献率

主成分 z_i 的贡献率为 下面的公式中应为λ_i

$$r_i / \sum_{k=1}^{p} \gamma_k (i = 1, 2, \cdots, p)$$

累计贡献率为

$$\sum_{k=1}^{m} \gamma_k / \sum_{k=}^{p} \gamma_k$$

④计算主成分载荷

$l_{ij} = \sqrt{\lambda_i} * e_{ij}$

六.马尔可夫预测方法

1.几个概念

1)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 如在商品 销售预测中，有“畅销”、“一般”、“滞销”等状态；在农业收成预测中， 有“丰收”、“平收”、“欠收”等状态

(2)状态转移过程 在事件的发展过程中，从一种状态转变为另一种状 态，就称为状态转移。譬如，天气变化从“晴天”转变为“阴天”、从“阴 天”转变为“晴天”、从“晴天”转变为“晴天”

(3)马尔可夫过程 若每次状态的转移都只仅与前一时刻的状态有关、而 与过去的状态无关，或者说状态转移过程是无后效性的，则这样的状态转移 过程就称为马尔可夫过程。‘

2）状态转移概率与状态转移概率矩阵

(1).状态转移概率

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。根据条件概率的定义， 由状态 Ei 转为状态 Ej 的状态转移概率 P（Ei→Ej)就是条件概率 P（Ej/Ei)， 即

$$P\ (E_i \rightarrow E_j)=P\ (E_j/E_i)=P_{ij}$$

(2)状态转移矩阵

假定某一种被预测的事件有 E1，E2，…，En，共 n 个可能的状态。记 Pij 为从状态 Ei 转为状态 Ej 的状态转移概率，作矩阵

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

称为状态转移矩阵

(3)状态转移概率矩阵的计算

采用频率近似概率的思想进行计算

2.马尔可夫方法

状态概率$\pi_i(k)$表示事件在 初始(k=0)状态为已知的条件下，经过 k 次状态转移后，在第 k 个时刻处于状态 E_i 的概率 根据马尔可夫过程的无后效性及 Bayes 条件概率公式，有

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

若记行向量$\pi(k)=[\pi_1(k),\ \pi_2(k),\ \cdots,\ \pi n(k)]$ 第 k 个时刻状态概率向量

可以递推推导出

$$\begin{cases} \pi(1)=\pi(0)P \\ \pi(2)=\pi(1)P=\pi\ (0)\ P^2 \\ \qquad M \\ \pi(k)=\pi(k-1)P=\cdots=\pi(0)Pk \end{cases}$$

P 为状态概率矩阵

1）第 k 个时刻(时期)的状态预测

如果π(0)已知 可通过上面递推公式求出π(k)

2)终极状态概率预测

经过无穷多次状态转移后所得到的状态概率称为终极状态概率，或称平 衡状态概率。

记终极状态概率向量为π 其应该满足

$\pi=\pi P$

$0<=\pi_i<=1$

$$\sum_{i=1}^{n}\boldsymbol{\pi}_i=1$$

七.趋势面分析方法

趋势面分析是利用数学曲面模拟变量在空间上的分布及变化趋势的一种数学方法。它实质上是通过回归分析原理，运用最小二乘法拟合一个二维非线性函数，模拟变量在空间上的分布规律，展示变量在地域空间上的变化趋势。

1.一般原理

通常把实际的地理曲面分解为趋势面和剩余面两部分，前者反映地理要素的宏观分布规律，属子确定性因素作用的结果；而后者则对应于微观局域，是随机因素影响的结果。趋势面分析的一个基本要求，就是所选择的趋势面模型应该是剩余值最小，而趋势值最大。

1)趋势面模型的建立

用来计算趋势面的数学方程式有多项式函数和傅里叶级数，其中最常用的是多项式函数形式多项式趋势面的形式为：

一次趋势面模型：$z=a_0+a_1x+a_2y$

二次趋势面模型：$z=a_0+a_1x+a_2y+a_3x^2+a_4xy+a_5y^2$

在实际的空间趋势面模拟中，按照对事物认识由易到难的规律，应首先考虑用一次趋势面模型的倾斜平面去拟合，然后再用二次抛物趋势面去模拟，如果还不能满足研究需求，则需选用三次趋势面、四次趋势面甚至更高次趋势面进行拟合。

2）参数估计

趋势面参数的估计 就是根据观测值 z_i,x_i,y_i 确定多项式的系数 $a_0,a_1\cdots a_p$

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

则 $\hat{z}=a_0+a_1x_1+a_2x_2+\cdots+a_px_p$

$$X=\begin{bmatrix}1 & x_{11} & x_{21} & \cdots & x_{p1}\\ 1 & x_{12} & x_{22} & \cdots & x_{p2}\\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn}\end{bmatrix}$$

$$A=\begin{bmatrix}a_0\\ a_1\\ \vdots\\ a_p\end{bmatrix}\quad Z=\begin{bmatrix}z_0\\ z_1\\ \vdots\\ z_n\end{bmatrix}$$

若 可以得到 $A=(X^TX)^{-1}X^TZ$

2.趋势面模型的适度检验

1）趋势面拟合适度的 R^2 检验

一般用变量 z 的总离差平方和中的回归平方和所占的比重表示回归模型的拟合优度

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

SS_D 是剩余平方和　　SS_R 回归平方和

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_D}{SS_T}$$

R^2 越大，趋势面的拟合度就越高

2）趋势面拟合适度的显著性 F 检验

$$F = \frac{SS_R / p}{SS_D / n - p - 1}$$

p 为自变量数

n 为样本数

3）趋势面适度的逐次检验

在多项式趋势面分析和检验中，有时需要对相继两个阶次趋势面模型的适度性进行比较，为此需要求出**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，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，然后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。若所得的 F 值是显著的，则较高次多项式对回归作出了新贡献，若 F 值不显著，则较高次多项式对于回归并无新贡献。

表 3.7.1　多项式趋势面由 K 次增高至 $(K+1)$ 次的回归显著性检验

离差来源	平方和	自由度	均方差	F 检验
$(K+1)$ 次回归	$SS_R^{(K+1)}$	p	$MS_R^{(K+1)} = SS_R^{(K+1)}/p$	$MS_R^{(K+1)}/MS_D^{(K+1)}$
$(K+1)$ 次剩余	$SS_D^{(K+1)}$	$n-p-1$	$MS_D^{(K+1)} = SS_D^{(K+1)}/(n-p-1)$	
K 次回归	$SS_R^{(K)}$	q	$MS_R^{(K)} = SS_R^{(K)}/q$	$MS_R^{(K)}/MS_D^{(K)}$
K 次剩余	$SS_D^{(K)}$	$n-q-1$	$MS_D^{(K)} = SS_D^{(K)}/(n-q-1)$	
由 K 次增高至 $(K+1)$ 次的回归	$SS_R^{(J)} = SS_D^{(K+1)} - SS_D^{(K)}$	$p-q$	$MS_R^{(J)} = SS_R^{(J)}/(p-q)$	$MS_R^{(J)}/MS_D^{(K+1)}$
总离差	SS_T			

用来判断高次趋势面是否对回归有了新贡献

八.地统计分析方法

一般认为 地统计学以区域化变量理论为基础，以变异函数为主要工具

1.区域化变量

区域化变量定义为以空间点 x 的三个直角坐标 x_u,x_v,x_w 为自变量的随机场 Z

也就是说区域化随机变量 是普通随机变量在一个域内 的确定位置上的特定取值

区域化随机变量具有两方面的含义，即观测前 Z(x)是一个随机场，观测后是一个普通的空间三元函数值。

比如在地理上去测量某个位置的湿度值，在测量前该位置的湿度值是未知的，是一个随机场，测量后为一个普通的空间三元函数值。

2.协方差函数

1）协方差函数的概念

在概率论中，随机向量 X 与 Y 的协方差被定义为

Cov(X,Y)=E[(X-EX)(Y-EY)]

区域化随机变量 Z(x)在空间点 x 和 x+h 处的两个随机变量 Z(x)和 Z(x+h)的二阶混合中心距定义为 Z(x)的自协方差函数

$$Cov[Z(x),Z(x+h)]=E[Z(x)Z(x+h)]-E[Z(x)]E[Z(x+h)]$$

2）协方差函数的计算公式

设 Z(x)为区域化随机变量，并满足 二阶平稳假设：随机函数 Z(x)的空间分布规律不因位移而改变

$$c^{*}(h)=\frac{1}{N(h)}\sum_{i=1}^{N(h)}[Z(x_i)-\bar{Z}(x_i)][Z(x_i+h)-\bar{Z}(x_i+h)]$$

$Z(x_i)$为在 x_i 处的实测值　N(h)为分割距离为 h 时的样本点对总数

若认为$\bar{Z}(x_i)=\bar{Z}(x_i+h)=m$ 可将 上式改写成

$$c^{*}(h)=\frac{1}{N(h)}\sum_{i=1}^{N(h)}[Z(x_i)Z(x_i+h)]-m^2$$

m 为样本平均数

$$m=\frac{1}{N}\sum_{i=1}^{n}Z(x_i)$$

3.变异函数

当空间点 x 在一维 x 轴上变化时，区域化变量 Z（x）在点 x 和 x+h 处的值 Z（x）与 Z（x+h）差的方差的一半为区域化变量 Z（x）在 x 轴方向上的变异函数，记为γ（h）.

$$\gamma(x,h)=\frac{1}{2}E[Z(x)-Z(x+h)]^2$$

当变异函数与为位置 x 无关时，改写为

$$\gamma(h)=\frac{1}{2}E[Z(x)-Z(x+h)]^2$$

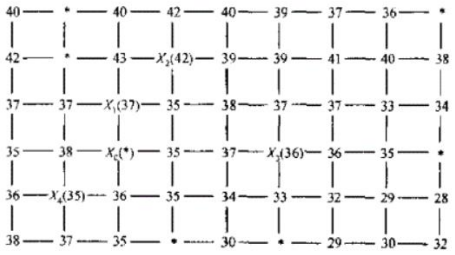
3）变异函数的计算公式

$$\gamma^*(h)=\frac{1}{2N(h)}\sum_{i=1}^{N(h)}[Z(x_i)-Z(x_i+h)]^2$$

对不同的空间分隔距离 h，根据协方差函数公式和变异函数公式就可以计算出相应的 c'（h）和γ（h）值来，再分别以 h 为横坐标，c”（h）或γ（h）为纵坐标，画出空间协方差函数曲线图和变异函数曲线图，可以直接展示区域化变量 Z（x）的空间变异特点。

举个例子

设某地区降水量Z（x）（单位：mm）是二维区域化随机变量，满已二阶平稳假设，其观测值的空间正方形网格数据如图所示点与点之间的距离为h=1km）。试计算其南北方向的变异函数。



对于无数据的点进行跳过

$$\begin{aligned}\gamma^*(1)=&\frac{1}{2\times36}[(40-42)^2+(42-37)^2+(37-35)^2+(35-36)^2+\\&(36-38)^2+(37-38)^2+(38-35)^2+(35-37)^2+(40-43)^2+\\&(43-37)^2+(36-35)^2+(42-42)^2+(42-35)^2+(35-35)^2+\\&(35-35)^2+(40-39)^2+(39-38)^2+(38-37)^2+(37-34)^2+\\&(34-30)^2+(39-39)^2+(39-37)^2+(37-36)^2+(36-33)^2+\\&(37-41)^2+(41-37)^2+(37-36)^2+(36-32)^2+(32-29)^2+\\&(36-40)^2+(40-33)^2+(33-35)^2+(35-29)^2+(29-30)^2+\\&(38-34)^2+(28-32)^2]\\=&385/72\approx5.35\end{aligned}$$

其他 h 情况类似计算

4)变异函数的参数

变异函数有四个非常重要的参数

基台值 变程 块金值 分维数

前 3 个参数可以直接从变异函数图中得到

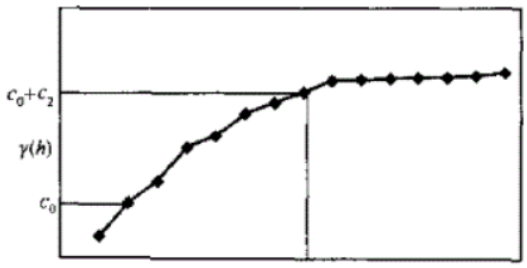


图 3.8.3 变异函数及有关参数

当变异函数值达到相对稳定的常数时称为基台值　达到基台值的间隔 a 称为变程　　当 h=0 时γ（0）值称为块金值
分维数由变异函数γ（h）和间隔距离 h 之间的关系决定

$$2\gamma(h) = h^{(4-2D)}$$

分维数 D 的大小表示函数曲线的曲率
5）变异函数的理论模型
①纯块金效应模型
②球状模型
③指数模型
④高斯模型
⑤幂函数模型
⑥对数模型
⑦线性有基台值模型
⑧线性无基台值模型

4.克里格法简介

克立格法是根据待估样本点（或块段）有限邻域内若干已测定的样本点数据，在考虑了样本点的形状、大小和空间相互位置关系，与待估样本点的相互空间位置关系，以及变异函数提供的结构信息之后，对待估样本点值进行的一种线性无偏最优估计。
2）克里格估计量
假设 x 是所研究区域内任一点，Z（x）是该点的测量值，在所研究的区城内总共有 n 个实测点，即 x_1，x_2…，x_n。那么，对于任意待估点或待估块段 V 的实际值 Zv（x），其估计值 Z;（x）是通过该待估点或待估块段影响范围内的 n 个有效样本值 Zv（x_i）（i=1，2，…，n）的线性组合来表示，即

$$Z_V^*(x) = \sum_{i=1}^{n} \lambda_i Z(x_i)$$

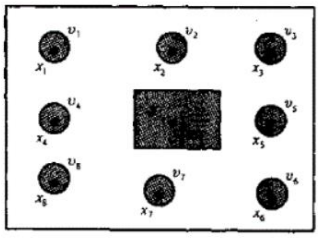


图 3.8.4　待估块段 V 与其邻域内的已知样本

克立格法的目标就是求一组权重系数λ_i（i=1，2…，n），使得加权平均值成为待估段 V 的平均值 Zv(x_0)的线性无偏最优估计量

令

$$K = \begin{bmatrix} \bar{c}_{11} & \bar{c}_{12} & \cdots & \bar{c}_{1n} & 1 \\ \bar{c}_{21} & \bar{c}_{22} & \cdots & \bar{c}_{2n} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ \bar{c}_{n1} & \bar{c}_{n2} & \cdots & \bar{c}_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}, \quad \lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ -\mu \end{bmatrix}, \quad D = \begin{bmatrix} \bar{c}(v_1, V) \\ \bar{c}(v_2, V) \\ \vdots \\ \bar{c}(v_n, V) \\ 1 \end{bmatrix}$$

则普通克里格方程组为

$$k\lambda=D$$

$$\lambda=K^{-1}D$$