笔记整理者：林鑫

[第一章：绪论](#第一章绪论)  
[第二章：地理数据及其采集与预处理](#第二章地理数据及其采集与预处理)  
[第三章：地理模型与地理建模概述（重点）](#第三章地理模型与地理建模概述重点）)  
 [SEM（结构关系模型）](#sem结构关系模型）)  
 [GWR（地理加权回归）](#gwr地理加权回归）)  
[第四章 地理学中的经典统计分析方法](#第四章-地理学中的经典统计分析方法)  
 [相关分析（初级）](#相关分析初级）)  
 [两要素之间](#两要素之间)  
 [秩相关系数（两种地理要素的分析方法）](#秩相关系数两种地理要素的分析方法）)  
 [线性相关计算及检验（两种地理要素的分析方法）](#线性相关计算及检验两种地理要素的分析方法）)  
 [多要素之间](#多要素之间)  
 [偏相关系数(多要素间相关程度)](#偏相关系数多要素间相关程度)  
 [回归分析（更高级的分析）先相关再回归](#回归分析更高级的分析）先相关再回归)  
 [线性回归](#线性回归)  
 [回归分析](#回归分析)  
 [自回归分析](#自回归分析)  
 [常见统计概念量](#常见统计概念量)  
 [残差平方和（RSS）](#残差平方和rss）)  
 [均方误差（MSE）](#均方误差mse）)  
 [均方根误差(RMSE)](#均方根误差rmse)  
 [平均绝对误差(MAE) :Mean Absolute Error](#平均绝对误差mae-mean-absolute-error)  
 [回归平方和（ESS）](#回归平方和ess）)  
 [误差平方和（SSE）](#误差平方和sse）)  
 [时间序列分析](#时间序列分析)  
 [时间序列的自相关性判断](#时间序列的自相关性判断)  
 [时间序列中的自回归分析](#时间序列中的自回归分析)  
 [季节变动的预测方法](#季节变动的预测方法)  
 [趋势面分析](#趋势面分析)  
 [空间分析与经典统计分析](#空间分析与经典统计分析)  
 [地理学第一定律（fig）](#地理学第一定律fig）)  
 [聚类分析](#聚类分析)  
 [直接聚类](#直接聚类)  
 [系统聚类分析](#系统聚类分析)  
 [K-Means均值聚类法](#k-means均值聚类法)  
 [主成分分析](#主成分分析)  
 [聚类分析与主成分分析两种方法如何组合使用？](#聚类分析与主成分分析两种方法如何组合使用)  
[第五章 空间统计分析初步](#第五章-空间统计分析初步)  
 [探索性空间统计分析](#探索性空间统计分析)  
 [经典分析在地学研究中的拓展](#经典分析在地学研究中的拓展)  
 [地理加权回归](#地理加权回归)  
 [权函数](#权函数)  
 [1.距离阈值法](#1距离阈值法)  
 [2.距离反比法](#2距离反比法)  
 [3.Gauss函数法](#3gauss函数法)  
 [带宽](#带宽)

第二章：地理数据及其采集与预处理

**第1节：地理数据的类型：**一、空间数据 二、属性数据

**第2节：地理数据的基本特征：**一、数量化、形式化与逻辑化 二、不确定性 三、多种时空尺度 四、多样性

**第3节：地理数据的采集与处理：**一、地理数据的采集 二、地理数据的处理

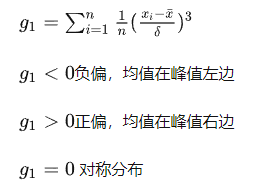
**第4节：地理数据的统计预处理：**基本步骤（一）统计分组（二）计算各组数据的频数、频率，编制统计分组表（三）作分布图

**二、几种常用的统计指标与参数：**（一）描述地理数据一般水平的指标：平均数，中位数，众数**，**（二）描述地理数据分布的离散程度的指标

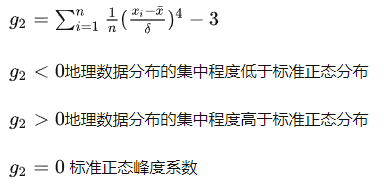
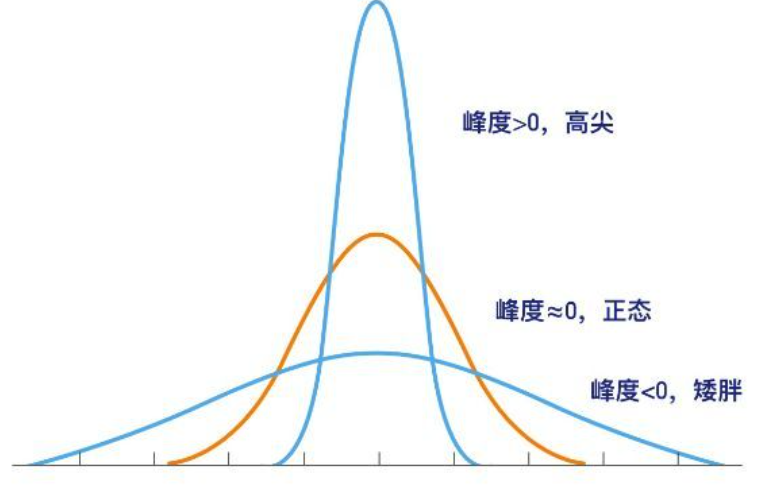
* 极差（所有数据中最大值与最小组之差）
* 离差（每一个地理数据与平均值的差）
* 离差平方和（总体上衡量一组地理数据与平均值的离散程度）
* 方差与标准差方
* 变异系数：表示了地理数据的相对变化（波动）程度

**（三）描述地理数据分布特征的参数**

**偏度系数：**测度了地理数据分布的不对称性情况，刻画了以平均值为中心的偏向情况



**峰度系数：**测度了地理数据在均值附近的集中情况



**第5节：地理数据分布的集中化与均衡度指数**

地理现象的分布格局，常常用地理数据分布的集中化程度与均衡度来描述

一、洛伦兹曲线与集中化指数

**洛伦兹曲线：**累积频率曲线研究工业化的集中化程度，洛伦兹曲线的上凸程度代表部门集中化程度 **集中化指数**，范围在[0,1]

二、基尼系数：洛伦兹曲线与集中化指数不能满足，对两组数据的进一步对比分析

三、锡尔系数：除了基尼系数外，也可以用基尼系数来进一步定量化描述数据

**第三章：地理模型与地理建模概述（重点）**

SEM（结构关系模型）Structural equation modeling

**基本概念——1.观测变量：**可直接测量的变量，通常是指标潜在变量无法直接观测并测量的变量。2**.潜变量：**需通过设计若干指标间接加以测量。3**.外生变量**：指那些在模型或系统中，只起解释变量作用的变量。它们在模型或系统中，只影响其他变量，而不受其他变量的影响。在路径图中，只有指向其他变量的箭头。没有箭头指向它的变量均为外生变量4.**内生变量：**是指那些在模型或系统中，受模型或系统中其它变量包括外生变量和内生变量影响的变量，即在路径图中,有箭头指向它的变量。它们也可以影响其它变量。

**具体结构：**结构方程模型由测量模型和结构模型构成**。** 测量模式旨在建立测量指针与潜在变项间的关系，主要透过验证性因素分析以检验测量指标(各种可实际观测的地理要素或指标)的效度。结构模式旨在检验潜在变项间之因果路径关系，主要针对潜在变项进行径路分析，以检验结构模式的适配性

第四章 地理学中的经典统计分析方法

**相关分析（初级）**

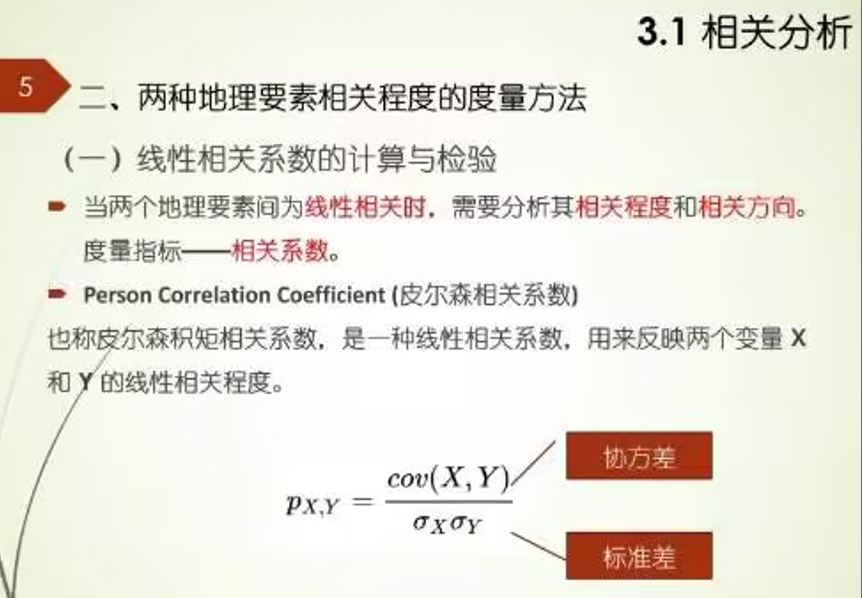
**两要素之间**

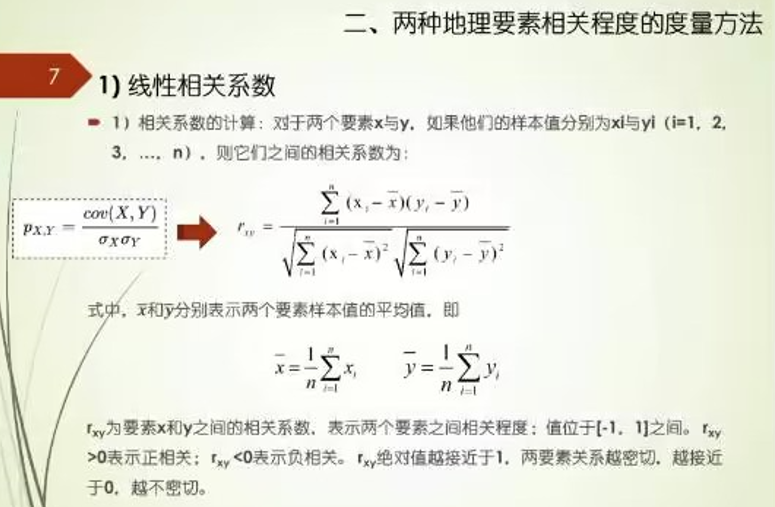
**秩相关系数（两种地理要素的分析方法）**，**基本概念：**秩相关系数（Coefficient of Rank Correlation），又称[等级相关系数](https://baike.baidu.com/item/等级相关系数/12730803?fromModule=lemma_inlink)，是将两要素的样本值按数据的大小顺序排列位次，以各[要素](https://baike.baidu.com/item/要素/5261200?fromModule=lemma_inlink)样本值的位次代替实际数据而求得的一种[统计量](https://baike.baidu.com/item/统计量/2112983?fromModule=lemma_inlink)。它是反映等级相关程度的统计分析指标，**常用的等级相关分析方法**有[Spearman相关系数](https://baike.baidu.com/item/Spearman相关系数/7977847?fromModule=lemma_inlink)和[Kendall秩相关系数](https://baike.baidu.com/item/Kendall秩相关系数/6246854?fromModule=lemma_inlink)等。

**线性相关计算及检验（两种地理要素的分析方法）**

地理模型可以分为线性模型和非线性模型

**计算**





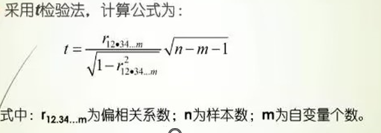
**多要素之间****：偏相关系数(多要素间相关程度)**

概念：在多要素构成的地理系统中，当研究某-个要素对另- 一个要素的影响或

相关程度时，其他要素的影响视为常量(保持不变)。即暂不考虑其他

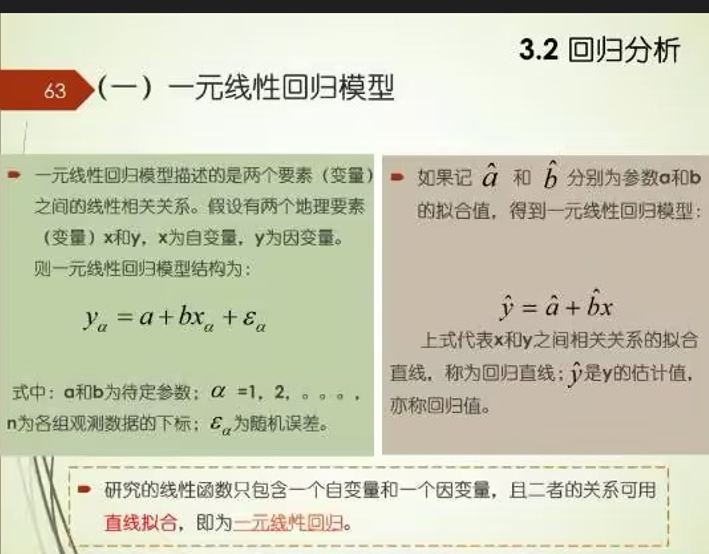
要素的影响，而单独研究两个要素之间的相互关系的密切程度时，则称

为偏相关。用以度量偏相关程度的统计量,称为偏相关系数。

原理：

**回归分析（更高级的分析）先相关再回归**

线性回归

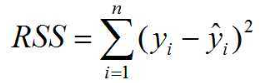


**自回归分析，为什么要进行？**

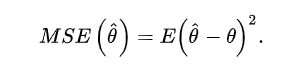
当一个要素(变量)按时间顺序排列的观察值之间具有依赖关系或相关性时，就可以建立该要素(变量)的自回归模型，并由此对其发展变化趋势进行预测。在建立自回归模型之前，首先需要对时间序列的自相关性进行判断。自相关性是建立自回归模型的基础，只有具有显著性的自相关性的时间序列才可以建立自回归模型。

**常见统计概念量**

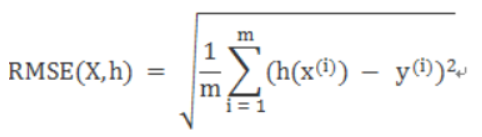
**残差平方和（RSS），**残就是拟合出来的比较，统计学上把数据点与它在回归直线上相应位置的差异称为残差，把每个残差平方之后加起来 称为[残差平方和](https://so.csdn.net/so/search?q=残差平方和&spm=1001.2101.3001.7020)（相当于实际值与预测值之间差的平方之和）。残差平方和是在[线性]模型中衡量模型拟合程度的一个量



**均方误差（MSE），**这里主要对象是参数，MSE是与参数估计中的比较**，**均方误差是指[参数估计](https://so.csdn.net/so/search?q=参数估计&spm=1001.2101.3001.7020)值与参数真值之差平方的期望值，在样本量一定时，评价一个点估计的好坏标准使用的指标总是点估计与参数[真值](https://so.csdn.net/so/search?q=真值&spm=1001.2101.3001.7020) 的距离的函数



**均方根误差(RMSE)：**均方根误差亦称标准误差，均方误差:均方根误差是均方误差的算术平方根

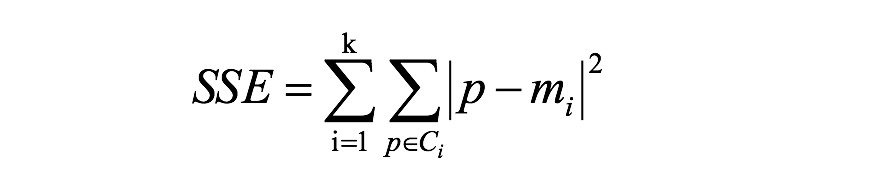


**平均绝对误差(MAE)** :Mean Absolute Error，平均绝对误差是绝对误差的平均值，平均绝对误差能更好地反映预测值误差的实际情况



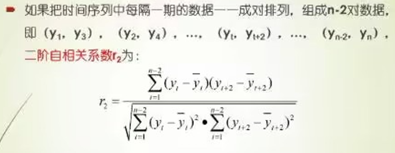
**回归平方和（ESS）：**因变量回归值ŷ-因变量平均值y的离差平方和，回归平方和ESS (Explained Sum of Squares)是[因变量](https://baike.baidu.com/item/因变量?fromModule=lemma_inlink)回归值ŷ-[因变量](https://baike.baidu.com/item/因变量?fromModule=lemma_inlink)平均值y的[离差平方和](https://baike.baidu.com/item/离差平方和/918673?fromModule=lemma_inlink)，数值上=∑(ŷ-ȳ)2，也称为解释平方和。用[回归方程](https://baike.baidu.com/item/回归方程/11042494?fromModule=lemma_inlink)或[回归线](https://baike.baidu.com/item/回归线/9443353?fromModule=lemma_inlink)来描述变量之间的统计关系时，实验值yi与按回归线预测的值ŷ并不一定完全一致。ESS越大说明多元线性回归线对样本观测值的拟合情况越好。

**误差平方和（SSE）：**误差平方和又称[残差平方和]、[组内平方和]等（Sum of the Squared Errors），根据n个观察值拟合适当的模型后，余下未能拟合部份(ei=yi一y平均)称为残差，其中y平均表示n个观察值的平均值，所有n个残差平方之和称误差平方和。在[回归分析](https://baike.baidu.com/item/回归分析/2625498?fromModule=lemma_inlink)中通常用SSE表示，其大小用来表明函数拟合的好坏。将残差平方和除以自由度n-p-1(其中p为自变量个数)可以作为误差方差σ2的[无偏估计](https://baike.baidu.com/item/无偏估计/3370664?fromModule=lemma_inlink)，通常用来检验拟合的模型是否显著。



**时间序列分析**

时间序列的自相关性判断，具体过程



**时间序列中的自回归分析**

当一个要素按时间顺序排列的观察值之间具有依赖关系或相关性时，就可以建立该要素的自回归模型，并由此对其发展变化趋势进行预测

建立模型前首先要进行时间序列的自相关性进行判断

自相关性是建立自回归模型的基础

**季节变动的预测方法**

时间序列的发展变化过程：长期趋势（T），季节变动（S），循环变动（C），不规则变动（R）

**三、季节变动的预测方法**

基本原理是用经过乘以李节系数的趋势线进行预测，基本步骤为:(1)将原时间序列求移动平均/滑动平均:(2)将原序列y除以其对应的趋势方程值(或平滑值)，即:季节系数=TSCI/趋势方程值(TC或平滑值) =SI(3)将月度(或季度)的季节指标加总，以由计算误差导致的值去除理论加总值，得到校正系数，并以该校正系数乘以季节性指标，调整后季节性指标。(4)进行预测:-下一年度的预测值， 延长趋势线即可各月(季)的预测值

**趋势面分析，趋势面有哪些模型**

定义：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  
本质是通过回归分析原理，运用最小二乘法（OLS）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

一般原理：趋势面是一种抽象的数学曲面，抽象并过滤掉了一些局域随机因素的影响，使地理要素的空间分布规律明显化**实际的地理曲面分解为趋势面和剩余面两部分**

（一）趋势面模型建立：计算趋势面的数学方程式有多项式函数和傅里叶级数

**趋势面模型的适度检验，**趋势面分析拟合程度与回归模型的效果直接相关

（一）趋势面拟合适度的检验（二）趋势面拟合适度的显著性F检验（三）趋势面适度的逐次检验

**空间分析与经典统计分析，**两大差异：空间数据并非独立，可能具有空间相关性

空间数据没有重复的，**具体区分**(1)空间数据间并非独立，而是在维空间中具有某种空间相关性，且在不同的空间分辨率下呈现不同之相关程度:(2)地球只有一个，大多数空间问题仅有一组(空间分布不规则的)观测值，而无重复观测数据。因此，空间现象的了解与描述是极为复杂的，而传统方法，尤其是建立在独立样本上的统计方法，不适合分析空间数据。

**地理学第一定律（fig）：**「地理学第一定律」：任何事物之间都是空间相关的，距离越近的事物之间的空间相关性越大。

**聚类分析：**聚类分析，也称为群分析或点群分析，是研究多要素事物分类问题的数量方法。

基本原理:根据样本自身的属性，用数学方法按照某种相似性或差异性指标，定量地确定样本之间的亲疏关系，并按这种亲疏关系程度对样本进行聚类。

常见方法有:系统聚类法、动态聚类法、模糊聚类法等。

**直接聚类****，**直接聚类的过程，需掌握过程。

**.直接聚类方法计算思维:**

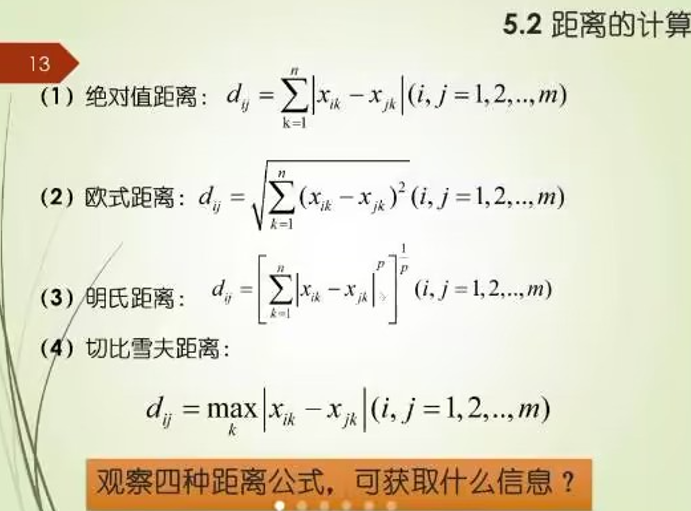
首先从距离矩阵中找到最小值，将最小值所在行和列的类合并，删除最小值所在行及行号对应的列，重复以上过程直至合并完所有类。如果其中一个分类对象已归于-类,则把另一个也归入该类。每-次归并，都划去该对象所在的行与行序相同的行，那么经过m-1次就可以把全部分类对象归为一类，这样就可以根据归并的先后顺序作出聚类谱系图。-根据距离矩阵结构依次并类得到结果，是-种简便的聚类方法。

**系统聚类分析，**又称群分析或点群分析，研究多要素事物分类问题的数量方法

**类类距离的选择：**最短距离法**，**最长距离法**，**重心法**，**组间平均距离法**，**离差平方和法

**类间距离法的选择：**Nearest neighbor(最短距离法)- -用两个类别中各个数据点之间最短的距离来表示两个类别之间的距离Furthest neighbor(最长距离法)-用两个类别中各个数据点之间最长的距离来表示两个类别之间的距离Centroid clustering(重心法)- -用两个类别的重心距离来表示两个类别之间的距离between-groups linkage(组间平均距离法)- -SPSS的默认方法。 是用两个类别中间各个数据点距离的平均来表示两个类别之间距离Ward's method(离差平方和法)- 使各类别中的离差平方和较小，而不同类别之间的离差平方和较大

**类内距离的选择**



**K-Means均值聚类法**原理，K-Mean算法，即 K 均值算法，是一种常见的聚类算法。算法会将·数据集分为 K 个第1步:确定要分的类别数目K。第2步:确定K类别的初始聚类中心。要求在用于聚类的全部样本中，选择K个样本作为K个类别的初始聚类中心使用SPSS进行聚类时，可以由系统自动指定初始聚类中心第3步:根据K个初始聚类中心，依次计算每个样本到K个聚类中心的欧氏距离，并根据距离最近原则将所有的样本分到事先确定的K个类别中第4步:计算K个类别每个变量的均值，并以均值点作为新的K个类别中心。根据新的中心位置，重新计算每个样本到新中心的距离，并重新进行分类。第5步:重复第4步，直到满足终止聚类条件:迭代次数达到研究者事先指定的最大迭代次数(SPSS默认10次)。新确定的聚类中心点与上一次迭代形成的中心点最大偏移量小于指定的重(SPSS默认0.02)。簇，每个簇使用簇内所有样本均值来表示，将该均值称为“质心”。

**第6节：#主成分分析**

一、主成分分析的基本原理;把原来多个变量划为少数几个综合指标的一种统计分析方法（数据降维技术）,即用较少的综合指标代替原来较多的变量指标，使较少的综合指标能较多反映原来信息（且独立）**最简单形式：原来变量的线性组合**

**原则:**减少数目，抓住主要矛盾，简化关系①独立——新变量相互无关②贡献——Z\_1Z1Z\_1Z1是一切线性组合中方差最大者（贡献最大，第一主成分）

二、主成分分析的计算步骤（1）计算相关系数矩阵（2）计算特征值与特征向量（3）计算主成分贡献率及累积贡献率（4）计算主成分载荷（也就是占比）

**聚类分析与主成分分析两种方法如何组合使用？**

1 主成分分析,主要思想在进行高维数据系统分析时，通过主成分分析，可以在纷繁的指标变量描述下，了解影响这个系统存在与发展的主要因素。主成分分析是研究如何通过少数几个主成分来解释多变量的方差的分析方法，也就是求出少数几个主成分，使他们尽可能多地保留原始变量的信息，且彼此不相关.

2 聚类分析,是主要思想聚类分析又称为群分析，是一种对指标或者样本进行分类的多元统计方法，是将数据分类到不同的簇当中，使得同一个簇之间的样本具有较大的相似程度，不同的簇之间相似度较低。

3 结合使用,SPSS主成分分析的结果是可以直接用来做聚类分析哈，做聚类分析是需要将数据归一化处理的，以保证数据的可比性。

第五章 空间统计分析初步

**探索性空间统计分析****,**ESDA（探索性空间数据分析）,explore spatial data analysis

一、探索性空间统计分析的基本原理

***（一）空间权重矩阵:***为了揭示现象之间的空间联系，首先需要定义空间对象的相互邻接关系,**空间权重矩阵是空间统计分析与经典统计分析的重要区别之一,**定义了一个空间权重矩阵W，描述了n个个体之间的空间依赖关系。

**建立空间权重矩阵的规则,**规则有很多，以下为常用的两种①简单的二进制邻接矩阵②基于距离的二进制空间权重矩阵

**空间关系概念化**

**概念**——理解:Conceptualization of Spatial Relationships空间关系的形象化描述信息。通俗的说，就是在进行分析之前，需要对你的空间关系，进行一个定义。

常见的空间关系概念化包括了距离、时间、区域、邻近、邻接等，具体使用哪个，取决于要测量和分析的对象是什么。

**种类****1.**反距离,即：随着距离的增大，影响就会减少。

通常使用距离这个概念来描述空间关系的话进行分析的话，还会去设定一个距离范围，或者说阈值，这样可以减少所需要的计算数（特别是对于大型数据集而言，至关重要）。2.距离范围,就是在同等距离范围内的所有要素，都认为每个事件对我的影响是同等的。3.无差别区域,其实就是把“反距离”和“固定距离范围”两个模型合二为一了。

**（二）全局空间自相关**

Moran指数IIII和Geary系数CCCC是两个用来度量空间自相关的全局指标,Moran指数反映的是空间邻接或空间邻接的区域单元属性值的相似程度，而Geary系数与Moran指数存在负相关关系,取值一般为-1~1

**Geary系数CCCC,**取值一般为0~2

**（三）局部空间自相关**

Moran指数IIII和Geary系数CCCC对空间自相关做了全局评估，但忽略了空间过程的潜在不稳定性问题,如果进一步考虑是否存在观测值的高低或低值的局部空间聚集等问题就需要考虑局部空间自相关

**LISA（空间联系的局部指标）**local indicators of spatial association

LISA考虑到了全局指标moran指数的分解，分解到每个值的贡献(局部moran)

**分类：** 局部Moran指数,局部Geary指数，**G统计**

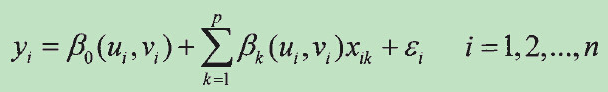
**Moran散点图；**研究局部的空间不稳定性，它是对空间滞后因子W\_zWzW\_zWz和z数据可视化的二维图示具有四个象界,分别对应于区域单元与其邻居直接的四种类型的局部空间联系形式

**经典分析在地学研究中的拓展**

**SEM和GWR:**结构方程模型(struc tural equation modeling, SEM) :研究地理要素因果关系、路径分析、因子分析、方差分析的综合手段  
地理加权回归(Geographical weighted regression,GWR) :研究地理要素空间可视关系、空间模式，空间异质性规律的有效手段

**地理加权回归****，**定义；[地理加权回归](https://baike.baidu.com/item/地理加权回归/5129442?fromModule=lemma_inlink)(Geographically weighted regression, GWR)是一种空间分析技术，广泛应用于地理学及涉及空间模式分析的相关学科。GWR通过建立空间范围内每个点处的局部回归方程，来探索研究对象在某一尺度下的空间变化及相关驱动因素，并可用于对未来结果的预测**，基本原理：**传统的ols（最小二乘模型）里面式子没有考虑到空间关系，而GWR考虑到了地理位置的影响（即加入了关于地理位置的函数）

GWR具体形式如下



**而是关键，它的估算就要用到权函数**

**关于权函数，带宽**

权函数：GWR模型的核心是空间权重矩阵，其选取的适当与否对回归参数的正确估计至关重要，比较常见的空间权函数有：1.距离阈值法；2.距离反比法；3.Gauss函数法等

1.距离阈值法，距离阈值法是最简单的权函数选取方法，其关键是选取合适的距离阈值D与dij进行对比，若大于阈值则权重为0，否则为1，形式如下： 2.距离反比法，假定空间相近的地物比较远的地物具有更强的相关性，基本形式为： 3.Gauss函数法，基本思想是通过选取一个连续单调递减函数表示权重与距离之间的关系，以此来克服以上两种方法的缺点。函数形式如下：

**式中b是描述权重与距离之间函数关系的非负数衰减参数，称为带宽。带宽越大，权重随距离增加衰减得越慢，反之则权重衰减得越快。**

**带宽：**带宽是核函数中高斯函数法中的非负数衰减参数， 描述权重与距离之间函数关系的非负数衰减参数，称为带宽。常用的最优窗宽选取准则有·\*\*CV交叉验证确认方法以及AICc信息准则\*\*