

ESRI 中国（北京）培训中心

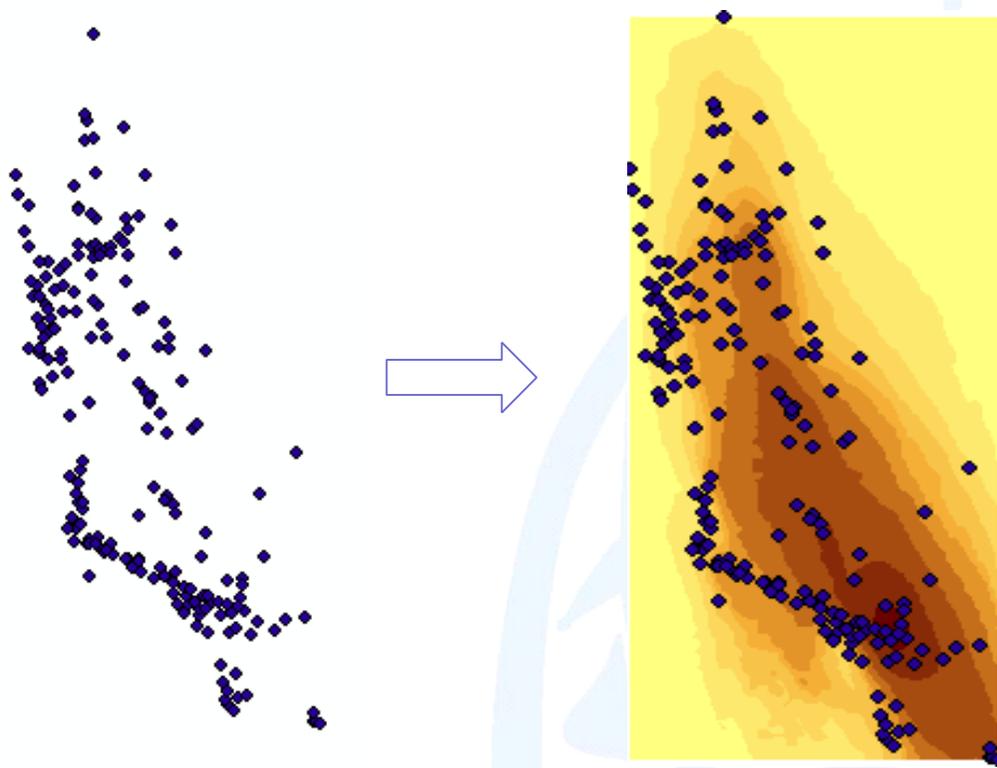
ESRI China(Beijing) Learning Center

探索性空间数据分析

王双 培训师

wangsh@lreis.ac.cn

楔子

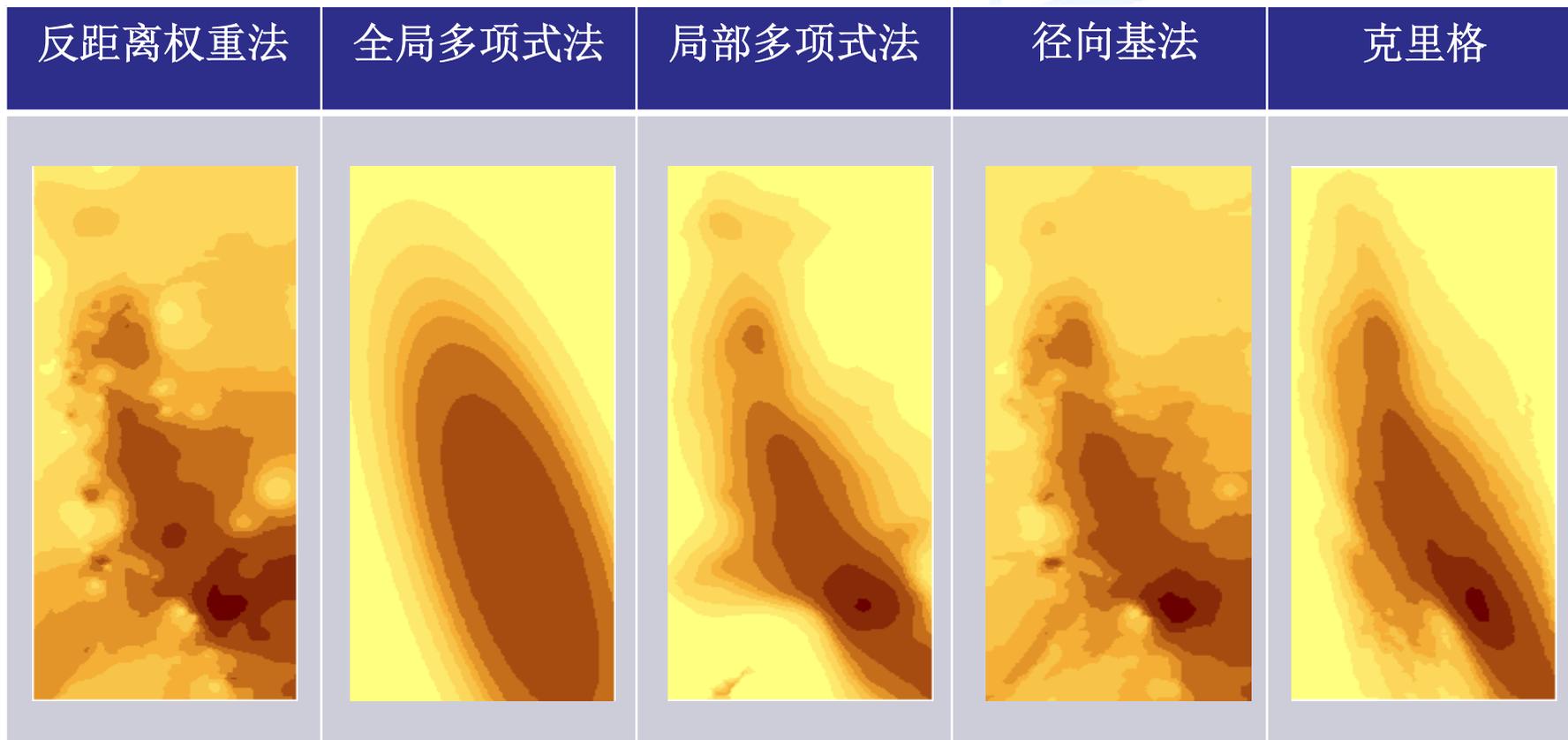


从离散样本点

得到连续表面

模型参数的设置
插值方法的选择

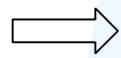
不同的插值方法



模型参数设置

- 有多少样本点参与到计算中来?
- 每个样本点的权重是相同的吗?
- 选择什么函数来模拟表面?
-

了解数据开始



探索性空间数据分析
Explore Spatial Data Analysis

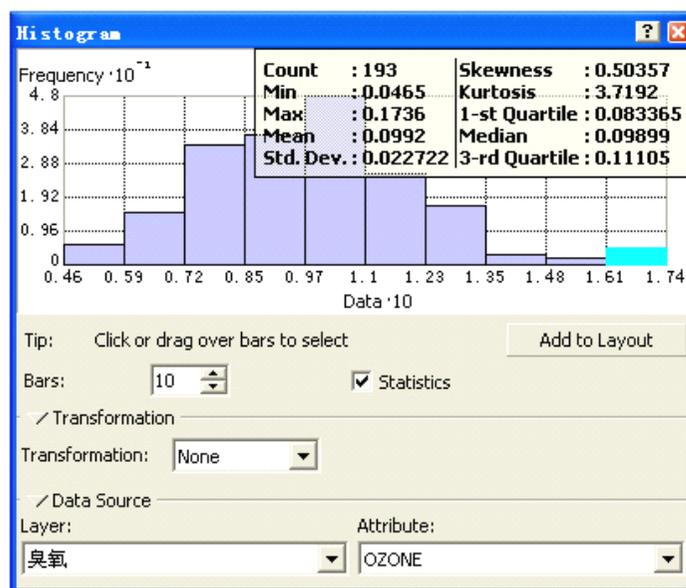
内容提纲

- 探索性空间数据分析
- 理解地统计
- 案例分析
- 模型比较



探索性空间数据分析

- 地理空间与属性空间
- 刷光（brushing）与链接（linking）



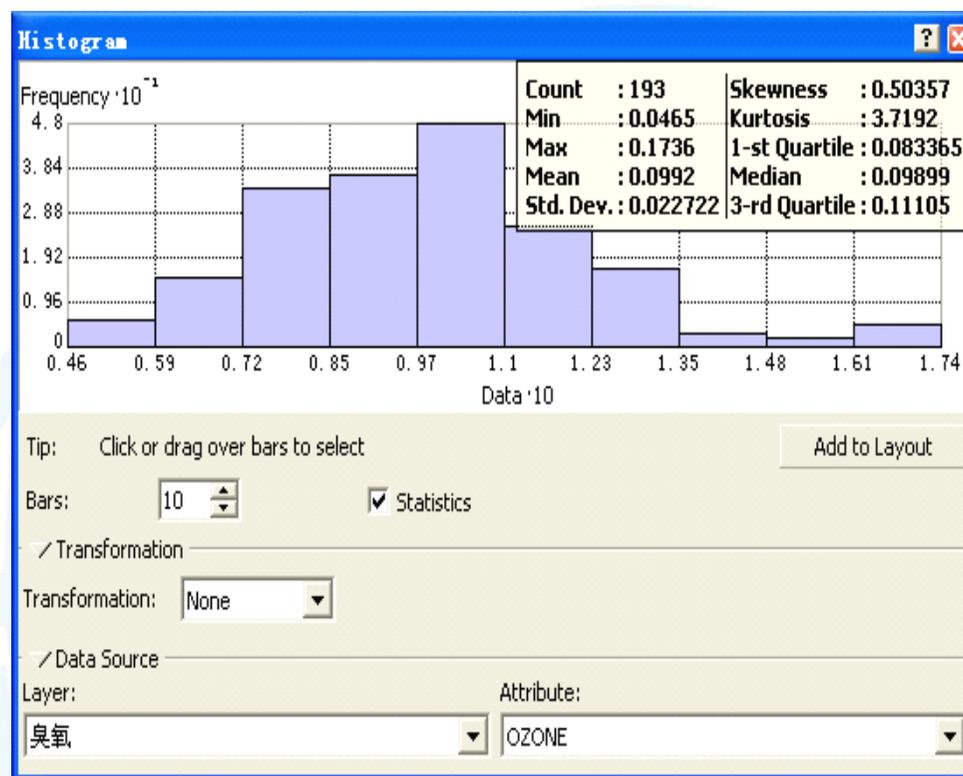
- 发现数据分布趋势以及离群值

探索性空间数据分析

- 直方图
- Normal QQ Plot
- 趋势分析
- Voronoi Map
- 半变异/协方差云图
- General QQ Plot
- 正交协方差云图

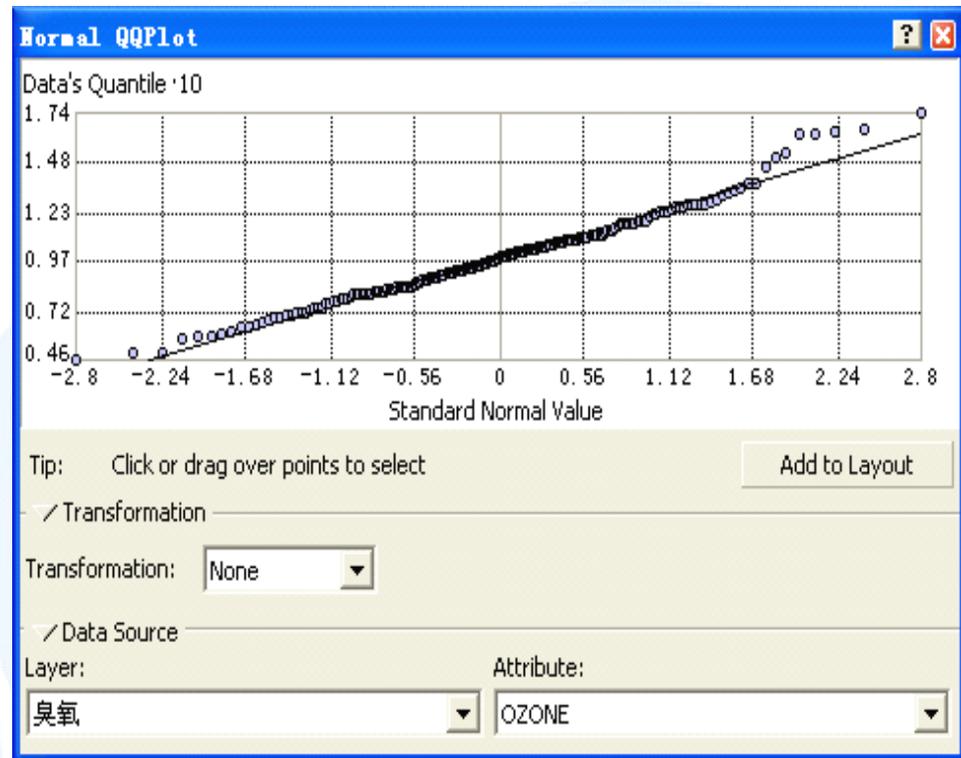
直方图

- 将数据分为若干区间，统计每个区间内的要素个数
- 给出一组统计量
- 检验数据是否符合正态分布以及发现离群值



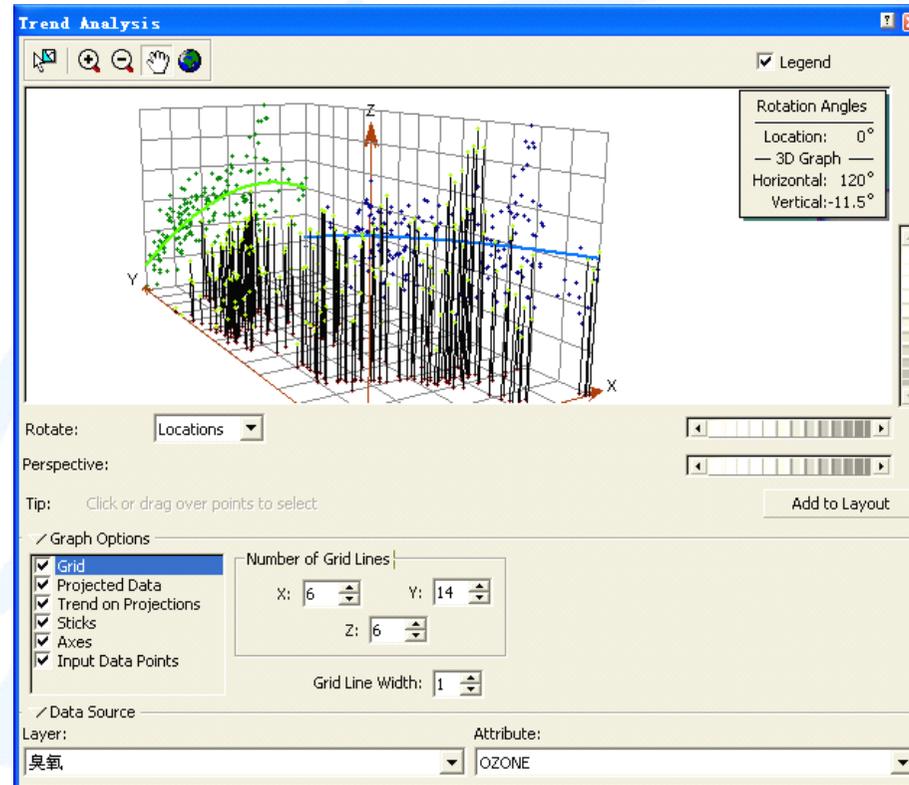
Normal QQ Plot

- 与标准正态分布（直线）对比
- 检验数据是否符合正态分布以及发现离群值



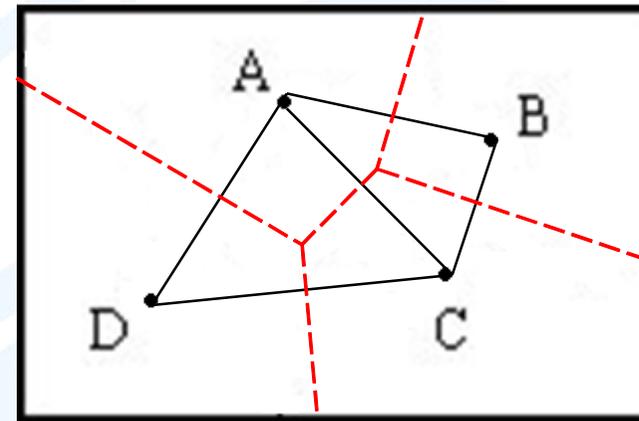
趋势分析

- 将每个采样点的值投影到东西方向和南北方向，发现数据在某个方向上有没有分布趋势



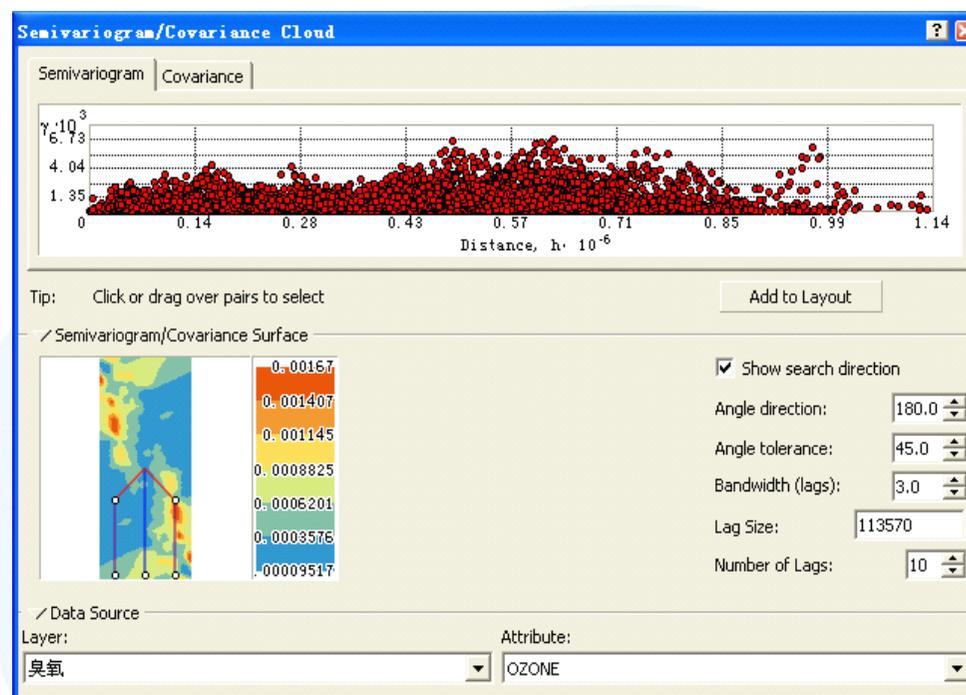
Voronoi Map

- 每个多边形内仅有一个样本点
- 多边形内的任一位置到该样本点距离最近



半变异/协方差云图

- 每一个点代表一个点对
- 空间距离越近，相关性越大
- 发现离群值以及是否存在各向异性



地统计方法

- 地统计 (Geostatistics) 又称地质统计, 它是以区域化变量为基础, 借助变异函数, 研究既具有**随机性**又具有**结构性**, 或空间相关性和依赖性的自然现象的一门科学。(汤国安, ArcGIS地理信息系统空间分析实验教程)
- 地统计核心: 根据样本点来确定研究对象 (某一变量) 随空间位置变化的规律, 以此来推算未知点的值。这个规律就是**变异函数**。

地统计方法

- 变异函数

基台值：反映最大变异情况，越高表明空间异质性越高；

变程：空间相关性的作用范围，超出变程则空间相关性不存在；

块金值：反映随机变化，受不确定性因素影响部分。

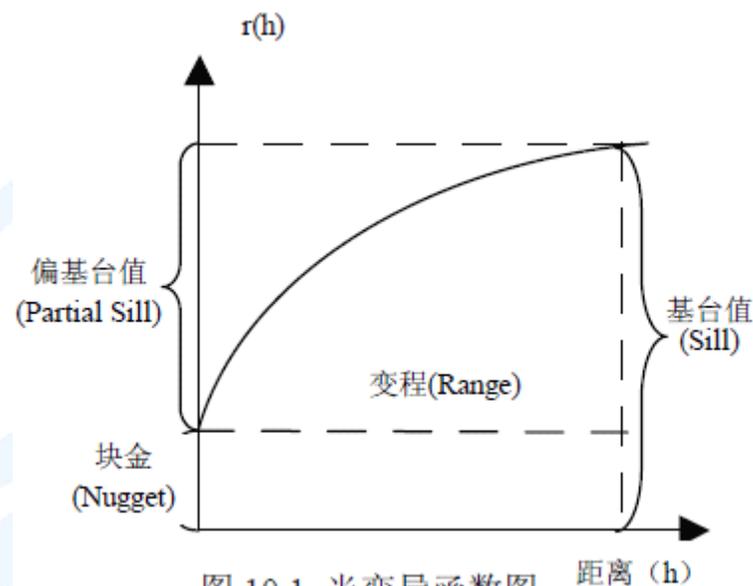
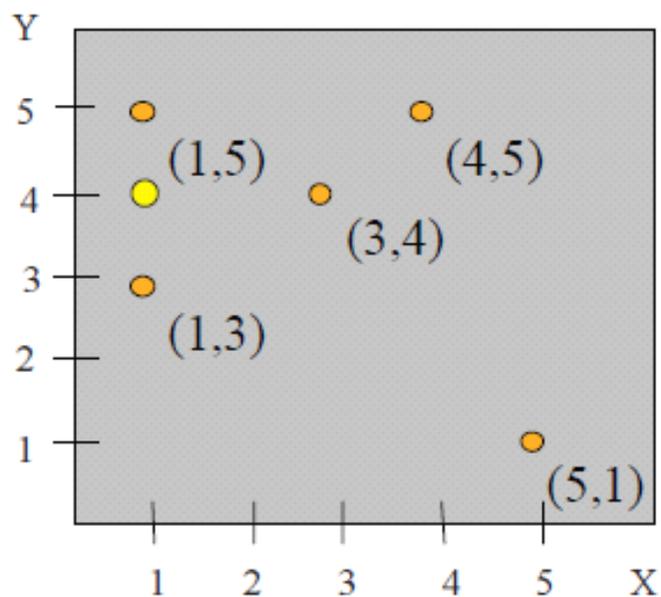


图 10.1 半变异函数图

地统计方法



样本点1	(1, 5)	100
样本点2	(3, 4)	105
样本点3	(1, 3)	105
样本点4	(4, 5)	100
样本点5	(5, 1)	115
预测点	(1, 4)	?

$$\hat{Z}(s_0) = \sum_{i=1}^N \lambda_i Z(s_i)$$

如何确定权重??

地统计方法

$$\mathbf{\Gamma} * \boldsymbol{\lambda} = \mathbf{g}$$

or

$$\begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1N} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma_{M1} & \cdots & \gamma_{MN} & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} * \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_N \\ m \end{pmatrix} = \begin{pmatrix} \gamma_{10} \\ \vdots \\ \gamma_{N0} \\ 1 \end{pmatrix}$$

γ_{ij} 表示两个样本点间基于距离的半变异值

\mathbf{g} 向量表示预测点与每个样本点间基于距离的半变异值

地统计方法

$$\text{半变异值} = 0.5 * [\text{第}i\text{点观测值} - \text{第}j\text{点观测值}]^2$$

表1 半变异值计算结果

位置	半变异值	距离计算	距离
(1,5) (3,4)	12.5	$\text{Sqrt}[(1-3)^2+(5-4)^2]$	2.236
(1,5) (1,3)	12.5	$\text{Sqrt}[(1-1)^2+(5-3)^2]$	2
(1,5) (4,5)	0	$\text{Sqrt}[(1-4)^2+(5-5)^2]$	3
(1,5) (5,1)	112.5	$\text{Sqrt}[(1-5)^2+(5-1)^2]$	5.657
(3,4) (1,3)	0	$\text{Sqrt}[(3-1)^2+(4-3)^2]$	2.236
(3,4) (4,5)	12.5	$\text{Sqrt}[(3-4)^2+(4-5)^2]$	1.414
(3,4) (5,1)	112.5	$\text{Sqrt}[(3-5)^2+(4-1)^2]$	3.606
(1,3) (4,5)	12.5	$\text{Sqrt}[(1-4)^2+(3-5)^2]$	3.606
(1,3) (5,1)	50	$\text{Sqrt}[(1-5)^2+(3-1)^2]$	4.472
(4,5) (5,1)	112.5	$\text{Sqrt}[(4-5)^2+(5-1)^2]$	4.123

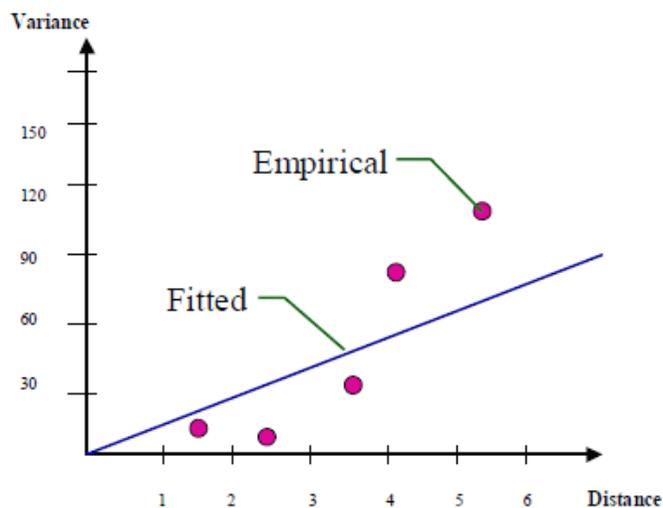
地统计方法

分组 (binning)

表2 理论半变异值的步长分组

步长间距	样点对间距	平均距离	半变异值	平均半变异值
1-2	1.414,2	1.707	12.5,12.5	12.5
2-3	2.236,2.236,3	2.491	12.5,0,0	4.167
3-4	3.606,3.606	3.606	50,12.5	31.25
4-5	4.472,4.123	4.298	50,112.5	81.25
>5	5.657	5.657	112.5	112.5

地统计方法



Variance=13.5*distance

表3 伽马矩阵 Γ 计算结果表

	(1,5)	(3,4)	(1,3)	(4,5)	(5,1)	
(1,5)	0	30.19	27	40.5	76.37	1
(3,4)	30.19	0	30.19	19.09	48.67	1
(1,3)	27	30.19	0	48.67	60.37	1
(4,5)	40.5	19.09	48.67	0	55.66	1
(5,1)	76.37	48.67	60.37	55.66	0	1
	1	1	1	1	1	0

地统计方法

表4 预测点 (1, 4) 的g向量计算结果

	距离	g向量
(1,5)	1	13.5
(3,4)	2	27
(1,3)	1	13.5
(4,5)	3.162	42.69
(5,1)	5	67.5
		1

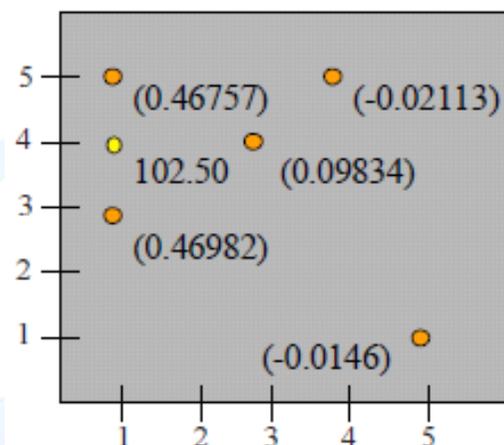


表5 预测结果

	权重	观测值	乘积
(1,5)	0.46757	100	46.757
(3,4)	0.09834	105	10.3257
(1,3)	0.46982	105	49.3311
(4,5)	-0.02113	100	-2.113
(5,1)	-0.0146	115	-1.679
(1,4)		预测结果	102.6218

案例分析（数据准备）

- 臭氧：1996年加州193个监测点臭氧浓度值的点要素，单位是ppm（parts per million）
- 城市：加州876个城市的点要素
- 目标：预测出每个城市的臭氧浓度值

案例分析（ESDA结论）

样本数	最小值	最大值	平均数	标准方差	偏态系数	峰态系数	中位数
193	0.0465	0.1736	0.0992	0.022722	0.50357	3.7192	0.09899

- 臭氧浓度基本呈正态分布，右偏分布，分布曲线比正态分布略陡
- 数据在不同方向上呈现不同的分布趋势（各向异性），在东北-西南方向呈现明显的倒U型分布（中间高两边低）
- 高浓度地区集中在洛杉矶附近，因为该地区人口稠密，车辆比较多

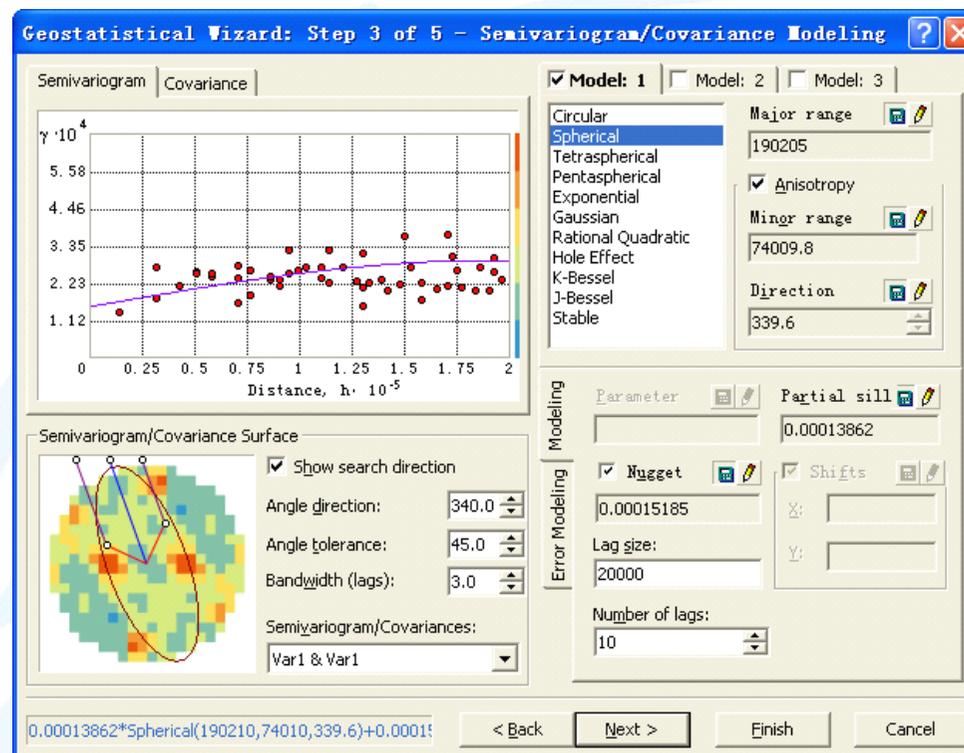
案例分析

- 选择插值方法
- 拟合变异函数
- 邻域分析
- 交叉验证
- 生成表面
- 对未知点估值

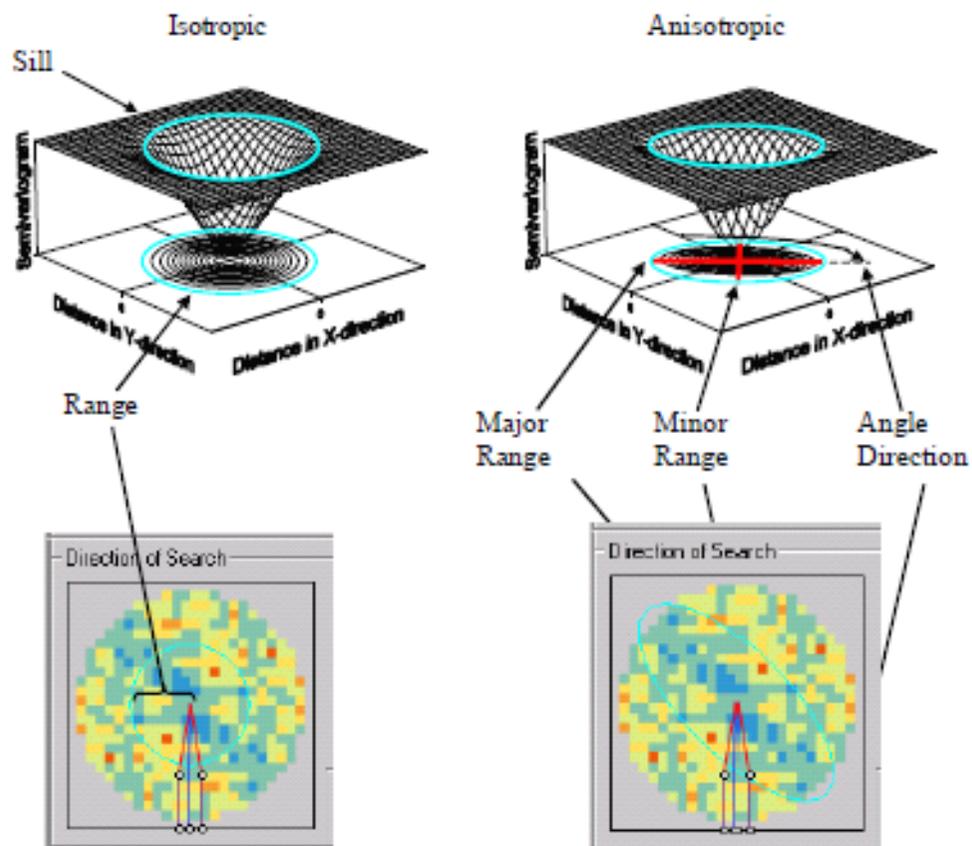


拟合变异函数

- 步长(lag size)和 number of lags
- 不同模型的选择
- 各向同性(Isotropic)与各向异性(Anisotropy)

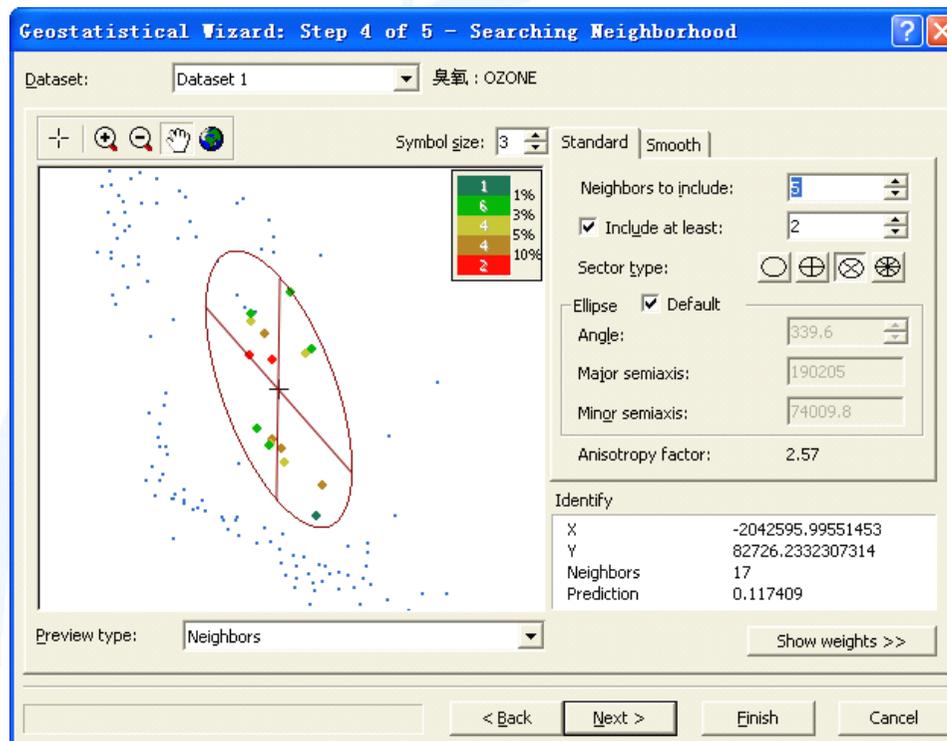


拟合变异函数



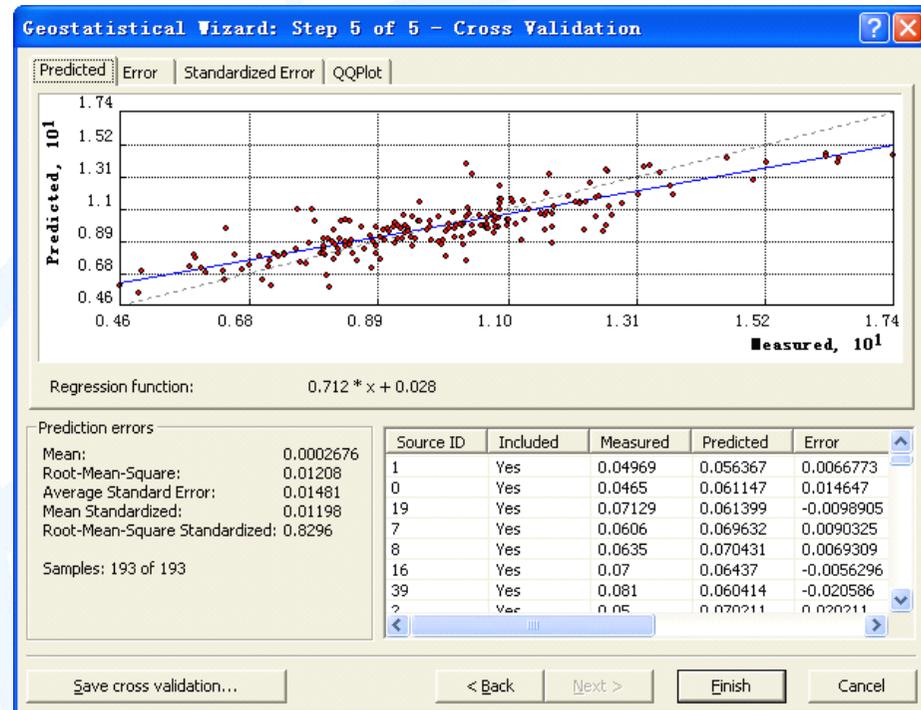
邻域分析

- 固定半径法
- 固定样本法（分区）



交叉验证

- 验证模型有多好
- 对每一个样本点给出模型的估计值，比较误差



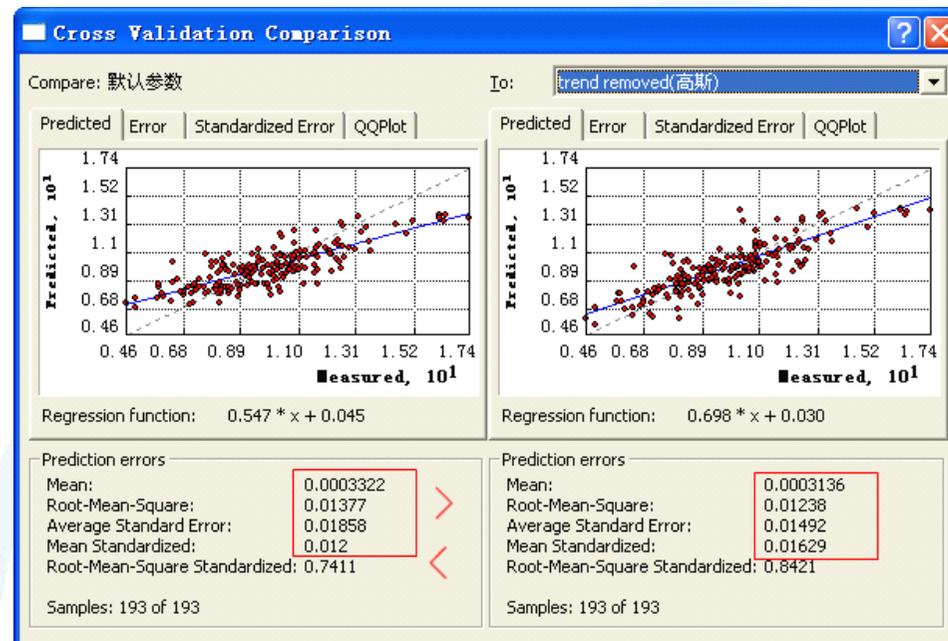
生成表面



ESRI 中国（北京）培训中心
ESRI China(Beijing) Learning Center

模型比较

- Mean Error: 平均误差
- Root mean square error: 均方根误差
- Average standard error : 平均标准误差
- Mean standardized error: 标准化平均误差
- Root mean square standardized error: 标准化均方根误差



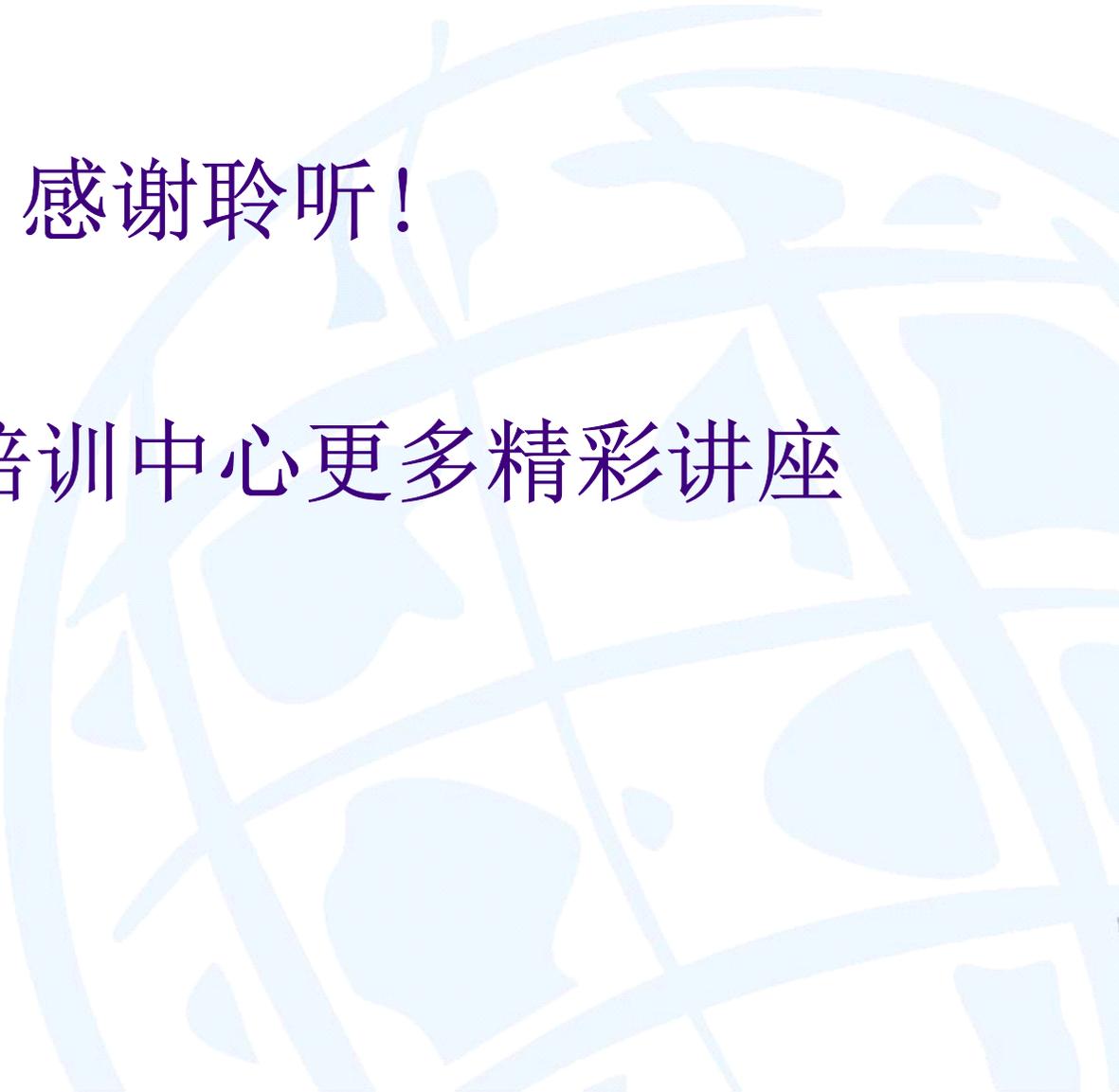
最优性：均方根误差较小

有效性：平均标准误差和均方根误差接近

小结

- 探索性空间数据分析
- 地统计与空间统计
- 地统计与空间分析





感谢聆听！

请关注培训中心更多精彩讲座