用 GeoDa 探索空间数据：工作手册

Luc Anselin

Spatial Analysis Laboratory

Department of Geography

University of ILLinois, Urbana-Champaign

Urbana, IL 61801

Http://sal.uiuc.edu/

Center for Spatially Integrated Social Science

Http://www.csiss.org/

Revised Version, March 6, 2005

Copyright © 2004-2005 Luc Anselin, All Right Reserved

翻译：郑光辉

2008 年 7 月 27 日星期日

前　言

本工作手册包括了一系列最初为 ICPSR 暑期空间分析课程空间数据分析和空间回归分析而准备的实验数据。它包括一系列的简要指南和经处理过的实例，这些实例作为 GeoDa™ User's Guide and GeoDa™ 0.95i Release Notes (Anselin 2003a, 2004)的补充。这些例子适合于 GeoDa™ 0.95i 版本，这一版本可以从网址 http://geoda.uiuc.edu/downloadin.php 免费下载。GeoDa 官方参考文献作者是 Anselin（2004c）。

GeoDa™ 是 Luc Anselin 所有的商标。

在这本手册中，有些材料中有些是在 SAL 网站可以下载的早期指南中已经有的。另外，还包括为一些课程而准备的数据，这些课程是在 2003 年秋季为 Illinois 大学农业与消费者经济系所开设，包括 ACE 492SA、空间分析和 ACE 492SE、空间计量经济学。由于 GeoDa 版本问题可能会存在一些小的差异。为避免疑问，当最新的文档取代了以前的指南材料时，通常是应该可以查阅到的。

这些实例和练习所用的样本数据都是可以从 SAL "stuff"网站上下载的。这些数据可以从以下网址下载 http://sal.uiuc.edu/data main.php。这些样本数据主要是为显示本软件的特征。建议读者只将它们作为练习之用。

声明：

这本手册的发展受益于美国国家科学基金（the U.S. National Science Foundation）通过 BCS9978058 项目给予空间综合社会科学（Center for Spatially Integrated Social Science ，CSISS）的支持。最近，又通过与疾病预防与控制中心（the Center for Disease Control and Prevention ，CDC）及预防医学教师联合会（Association of Teachers of Preventive Medicine ，ATPM）得到资助，award # TS-1125（不懂）。最后，许多 GeoDa 工作小组的成员们都提出过许多建议或意见，在此表示感谢。特别感谢 Julia Koschinsky，他仔细检查过每一个早期的版本，修改了许多错误之处。

练习 1　启动 GeoDa

1.1 目标

本次练习演示了如何开始使用 GeoDa 和它的用户界面的基本构成。练习完成之后，您应该掌握：

- 打开和关闭一个工程
- 载入一个带有指示器（主关键字）的 shape 文件
- 从菜单或工具栏中选择功能

有关这些操作的详细信息请参考 the User's Guide 第 3–18 页和 the Release Notes 第 7–8 页。

1.2 打开一个工程

双击桌面图标或通过 Windows Explorer（相应目录中）启动 GeoDa，就会出现一个欢迎界面。在 File 菜单中选择 Open Project，或单击 Open Project 工具按钮，如图 1.1 所示（on p. 2）。工具栏中只有两项是可用的，第一个是用于打开一个项目，如图所示。另一个项目是关闭项目（参见图 1.5（on p. 4））

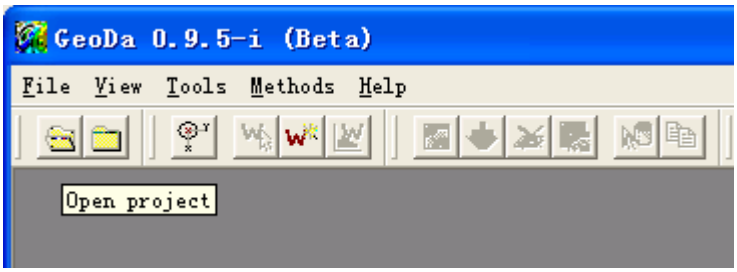


图 1.1　第一个菜单和工具栏

打开一个项目之后，熟悉的 Windows 对话窗口要求输入 shape 文件的文件名及可用的关键字。关键字是每一记录（observation）的唯一标识，它是一个整数值，如各个县的 FIPS 码，或人口（tract）数。

GeoDa 只能读入 shape 文件打开一个项目（at this point），但如果你的数据不是 shape 文件格式，可以用软件自带的空间数据处理工具来创建（请参考练习 4 和练习 5）。

在出现的文件对话框中选择 SIDS2 作为 Input Map，关键字默认为 FIPSNO。你也可以键入 shape 文件的完整路径，或在 Windows 文件结构中查找，直到出现文件名（对话框中只显示 shape 文件）。最后点击 OK，载入地图，如图 1.2。

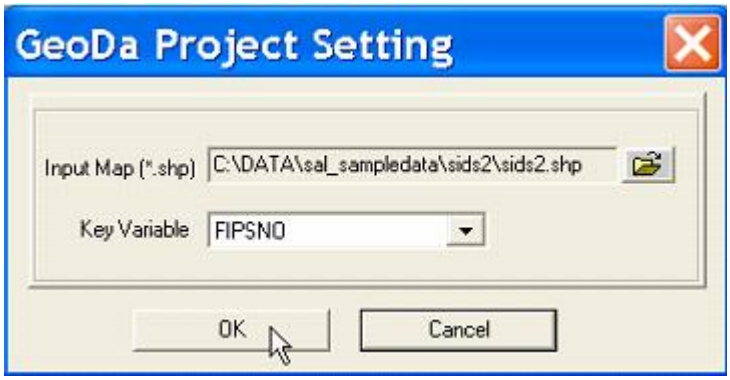


图 1.2　选择输入 shape 文件

接着，打开一个地图窗口，显示用于分析的底图，图中显示了 North Carolina 的 100 个县，如图 1.3 所示。窗口左侧区域显示了图例。通过向左或向右拖拽两区（图例和地图）之间的分隔线可以改变其大小。

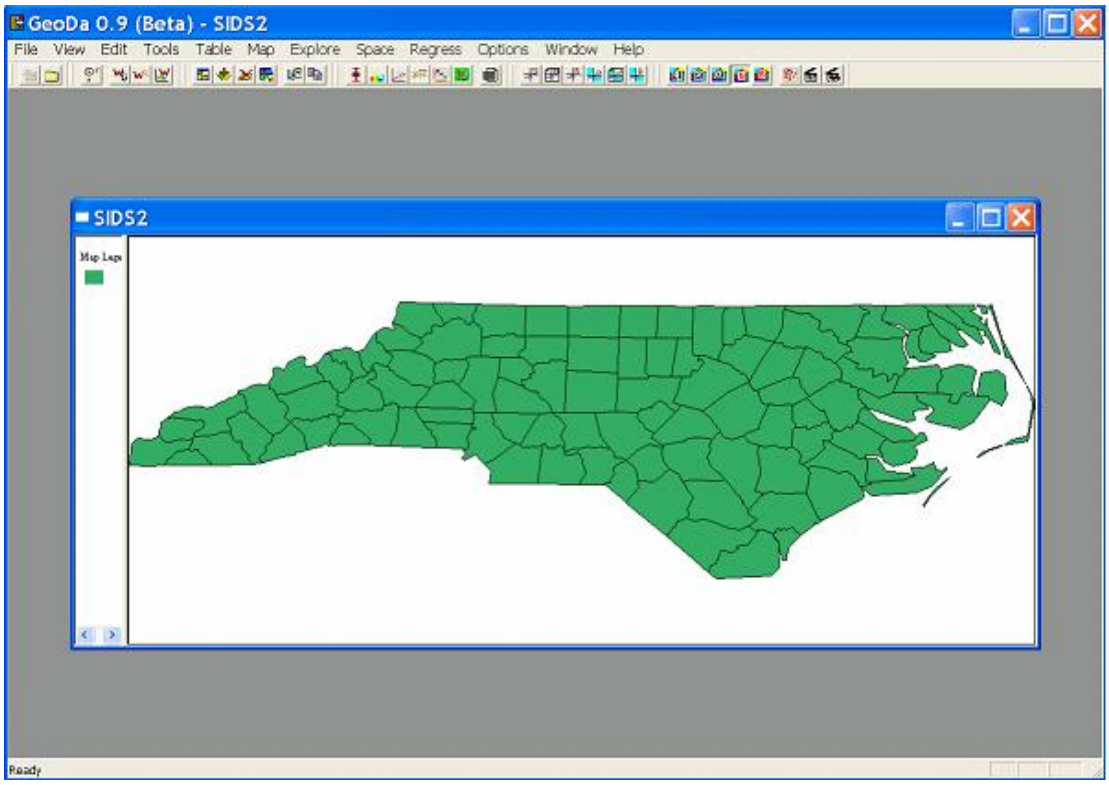


图 1.3　载入 SIDS2 样本数据后打开的窗口

你可以在地图窗口中左击选择颜色（背景，阴影等）属性来改变底图设置，及改变选择工具形状。左击打开一个菜单，如图 1.4 所示。例如，底图默认为绿色，将其改变为其它颜色，单击 Color>Map，从标准 Windows 颜色面板中选择一种新的颜色。

要关闭所有窗口，单击工具按钮中的 Close all windows（图 1.5 on p.4），或选择 File 菜单中的 Close All。

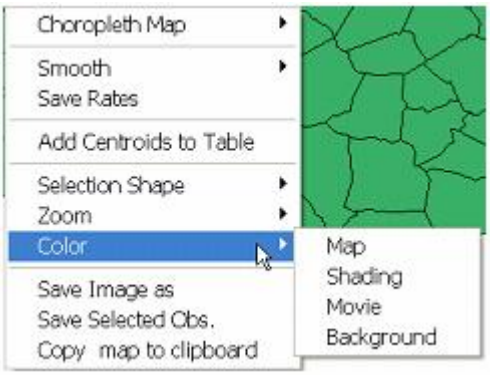


图 1.4　地图中的选项（左击）

图 1.5　关闭所有窗口

1.3 用户界面

载入一个 shape 文件之后，所有的菜单和工具栏都变为可用，详见图 1.6。

图 1.6　完整的菜单和工具栏

菜单栏由 11 项组成，4 项是标准 Windows 菜单：File（打开和关闭文件），View（选择要显示的工具栏），Windows（选择或重新排列窗口），和 help（还不能使用）。GeoDa 特有的有 Edit（控制地图窗口和图层），Tools（空间数据处理），Table（数据表格处理），Map（choropleth 制图和地图平滑（smoothing）），Explore（统计图表），Space（空间自相关分析），Regress（空间回归）和 Options（特殊（specific）应用选项）。你可以单击不同的菜单试一下 GeoDa 的功能。

工具栏由 6 组按钮组成，从左到右分别是：打开和关闭项目；空间权重计算；编辑功能；探索性数据分析；空间自相关；等级平滑和制图。以探索性工具栏为例，如图 1.7 所示。

图 1.7　探索工具

单击工具栏中按钮与在菜单中选择相应的选项功能是相同的。工具栏可以放置于任何位置，也就是说你可以将它们移动到任何不同位置。选择一个工具栏，（Experiment with this）单击左侧分割栏的鼓起处，把它拖到不同的位置。

1.4 练习

首先关闭所有 North Carolina 的数据图。利用 St. Louis 都市区周边 78 个县的杀人案数据（stl_hom.shp）打开一个新的项目，FIPSNO 作为关键字。试验一些地图选项，如底图颜色（Color>Map）

或窗口背景颜色（Color>Background）。在操作之前一定要关闭所有窗口（？）。

练习 2　建立一幅 Choropleth 图

2.1 目标

本练习演示一些制图及在地图中进行选择的所需的基本操作。

在练习之后，你将会学会：

- 制作一幅简单的 Choropleth 图
- 在地图中选择项目
- 改变选择工具

有关这些操作的更详细的信息请参考用户指导的 35－38 页及 42 页。

2.2 四分位图

样本集中的 SIDS 数据来自于 Noel Cressie's（1993）的空间数据统计学（Cressie 1993, pp. 386–389）。它包含了 North Carolina 的 100 个县两个年代 SIDS（婴儿猝死综合症，sudden infant death syndrome）死亡人数，分别表示为 SID74 和 SID79。另外也有每个州的出生人口数（BIR74，BIR79）和一个子数据集，还有非白种人出生人口数（NWBIR74，NWBIR79）。

首先利用练习 1 中的程序打开 sids.shp 文件。同前，选择 FIPSNOSNO 作为关键字。一幅 North Carolina 州各县的绿色底图展现于你的面前，如图 1.3 所示。唯一的不同就是窗口标题将会是 sids，而不是 SIDS2。

要新建两幅四分位图，比较非白种人在 1974 年（NWBIR74 和 SID74）出生人数与 SIDS 死亡的人数的空间分布。单击底图使其成为当前窗口（在 GeoDa 中，最后单击的窗口为当前窗口）。在 Map 菜单中选择 Quantile，将会出现一个对话框，以选择制图所用的变量。另外也会出现一个数据表。这个暂时不谈[1]。你应将其最小化，以防其碍事，但你稍后就会用到它，所以不要关闭[2]。在变量选择窗口，选 NWBIR74，如图 2.1 所示，然后单击 OK。注意选中对话框中的 check box，使所选变量成为默认选项。如果选中，下次就不会再要求选择变量了。如果你要对同一变量作几种不同类型的分析，这将会比较方便。但是，你本例中，我们要对不同变量进行相同的的分析，所以设置一个默认变量并不明智。如果你不小心选中了默认选项，你可以调用 Edit 菜单中 Select Variable 来取消。

[1] 第一次指定函数所需的变量，就会出现这一表格。
[2] 通过单击窗口右上角最左边的按钮使其最小化。

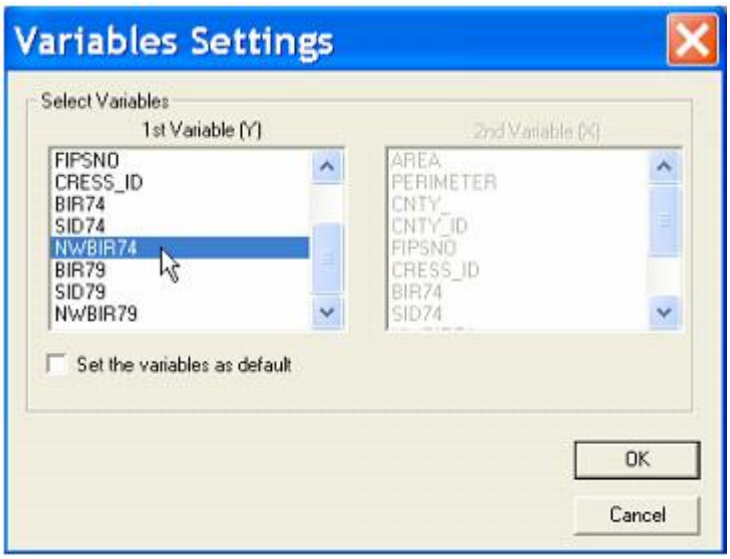


图 2.1　变量选择

选择变量之后，第二个对话框将要求输入四分位图的分类数量：暂时保留默认值 4（四分位图），点击 OK。就会出现一幅四分位图（4 类），如图 2.2 所示。

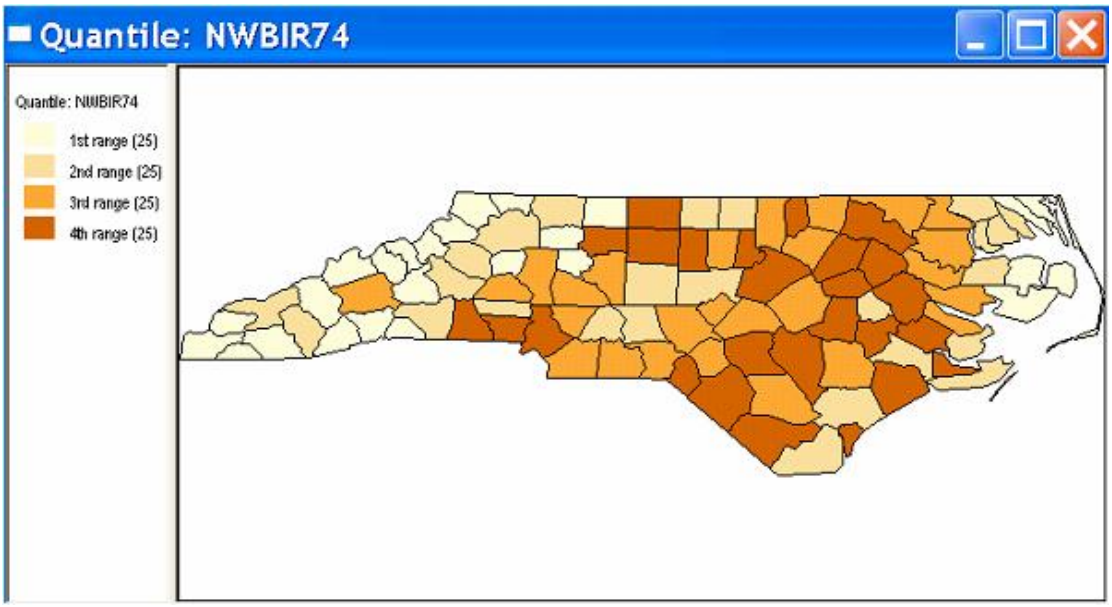


图 2.2：非白种人 1974 年出生人口数（NWBIR74）各县四分位图

利用同一底图再建一幅 choropleth 地图。首先，通过单击 Duplicate map 工具按钮再打开一个底图窗口，如图 2.3 所示。同样地，你也可以从菜单选择 Edit>Duplicate Map。

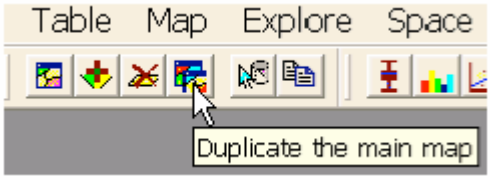


图 2.3　复制地图工具按钮

然后为变量 SID74 创建一幅四分位图（4 类），如图 2.4 所示）。你注意到每张四分位图中的观测数量有何变化？这贴图有两个问题。第一，这是一幅有关数量的 choropleth 图，或称为 *extensive* 变量，这种变量与大小（如面积或总人口）有关系，经常是不合适的。相反，级别或密度适合制作

choropleth 图，称为 *intensive* 变量。

第二是断点计算问题。对于如 SIDS 死亡数的分布问题，或多或少服从 Poisson 分布，低值之间存在许多联系。本例中断点的计算并不可靠，四分位图特别是五分位图是不正确的。注意最低类别显示为 0 个记录，第二类为 38 个。

你可以从菜单选择 Edit>Copy to Clipboard 将地图保存于剪切板。这只拷备了地图部分。如果你还希望得到图例的拷备，右击图例区域选择 Copy Legend to Clipboard。或者，你也可以从菜单中选择 File>Export>Capture to File，将其保存为.bmp 格式的地图（但无图例）。你需要指定一个文件名（如果需要也要指定路径）。如有需要，你可以利用图形转换软件包将 bmp 格式文件转换为其它格式。

2.3 在地图中选择和链接观测对象

到现在，这些地图还是“静态的”。动态地图的意思是说可以在不同地图之间选择某一地图，并链接所选择的地图。Geoda 包括许多形状选择工具，如点、矩形、多边形、圆形、直线。点状及多边形是多边形 shape 文件的默认选择工具，圆形是点状 shape 文件的默认选择工具。单击它的位置来选择一个对象（在某一县的位置上单击来选择它），或通过拖动鼠标来选择多个对象（在一个点上单击，拖动这一点到其它地方来创建一个多边形，然后松开。你可以通过 shift 键与单击来添加或去除被选择对象。地图外任何地方单击即可取消选择。其它的选择形状，可以在地图上右击，在 Selection Shape 下拉菜单中选择相应选项，如图 2.5 所示。每一幅地图都有其选择工具，它们彼此间不必相同。

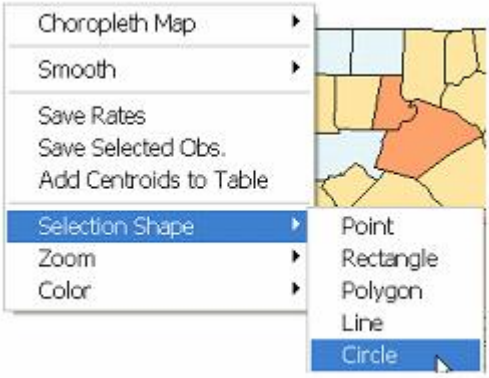


图 2.5　形状选择下拉菜单

比如，你选择圆形选择工具（如图 2.5），在 NWBIR74 中单击，扩大圆形边界来选择一些县（见图 2.6）。

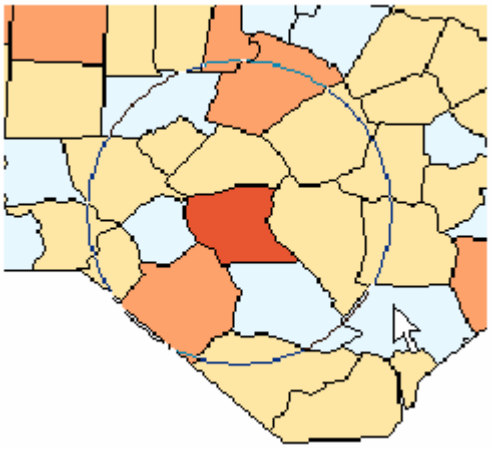


图 2.6　圆形选择

一旦松开鼠标，圆心在圆形区域内的县即被选中，显示为网状阴影（图 2.7）。当多幅地图同时在使用，所有地图中的相同县都会被选中，如图 2.7 两幅地图中网状阴影所示。这就是所谓的链接，这不仅适用于地图，也适用于打开的表格及其它静态图表。你可以通过地图选项改变网格的颜色（可击 Color>Shading）。

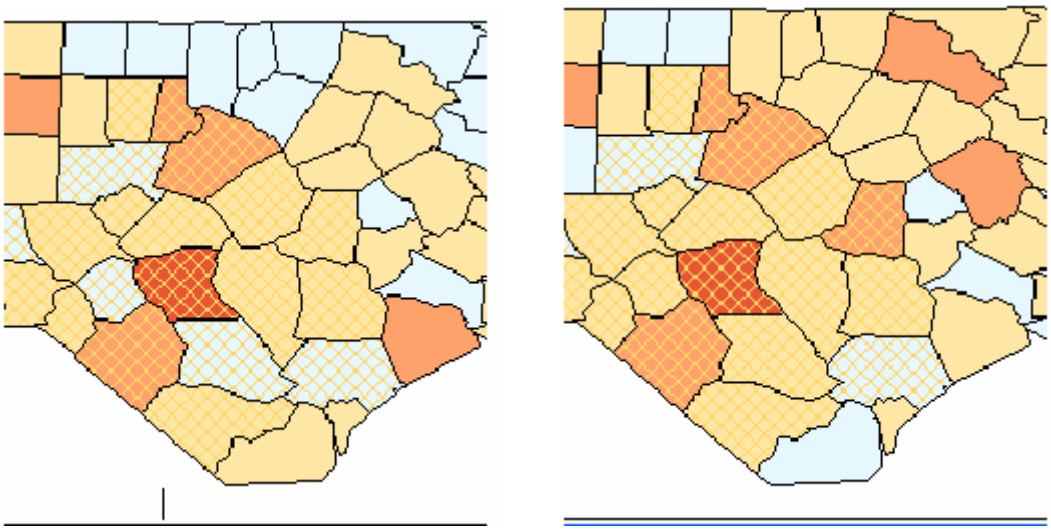


图 2.7　在链接地图中选择

2.4 练习

关闭所有窗口，利用 St.Louis 杀人案数据(stl_hom.shp，关键字 FIPSNO)打开一个新的项目。创建两幅 78 个县杀人案发生率的四分位图（5 类），一幅为 84 年－88 年（HR8488），一幅为 88 年－93 年（HR8893）。利用地图菜单及右击的方法创建 choropleth 图。利用不同的形状选择工具在其中一幅地图中选择一些县。注意在其它地图中相同的县也被选中。如果你愿意，可以将其中一幅保存为 bmp 文件，插入到 MS Word 文件中。试验第二种类型图，标准差图，它是以标准差单元值进行排序。

练习 3　基本表操作

3.1 目标

本练习显示了使用表格中的功能所需的一些基本操作，包括创建和转换变量。

在本练习后，你应学会：

- 打开和浏览（navigate）数据表
- 在表格中选择一项或排序
- 在表格中创建新的变量

有关这些操作的详细信息请参考用户手册。

3.2 浏览数据表

关闭所有窗口，载入 sids.shp 样本数据（以 FIPSNO 为关键字）。以一个变量（如 NWBIR74）创建一幅 choropleth 图，用选择工具选择一些县。如果先前将表格窗口最小化了，现在将其恢复到当前窗口。向下滚动表格，注意被选中的县为蓝色高亮显示，如图 3.1 所示。

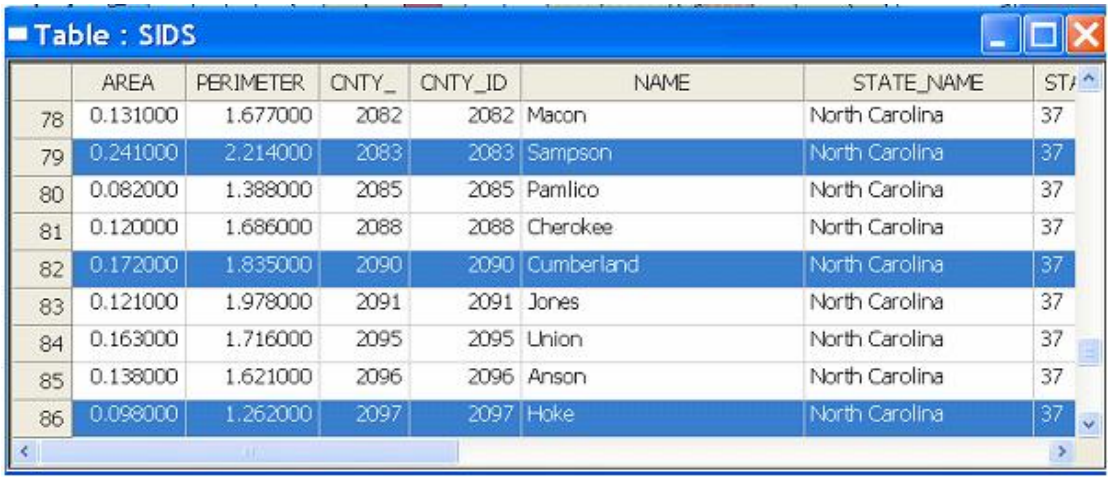


Table : SIDS

	AREA	PERIMETER	CNTY_	CNTY_ID	NAME	STATE_NAME	STA
78	0.131000	1.677000	2082	2082	Macon	North Carolina	37
79	0.241000	2.214000	2083	2083	Sampson	North Carolina	37
80	0.082000	1.388000	2085	2085	Pamlico	North Carolina	37
81	0.120000	1.686000	2088	2088	Cherokee	North Carolina	37
82	0.172000	1.835000	2090	2090	Cumberland	North Carolina	37
83	0.121000	1.978000	2091	2091	Jones	North Carolina	37
84	0.163000	1.716000	2095	2095	Union	North Carolina	37
85	0.138000	1.621000	2096	2096	Anson	North Carolina	37
86	0.098000	1.262000	2097	2097	Hoke	North Carolina	37

图 3.1　链接表中被选中的县

为了更容易识别被选中的地区（如查看被选中县的名字），可以利用 Table 菜单中 Promotion。可以通过 Table 下拉菜单来调用些命令（表格中任意位置右击），如图 3.2 所示。被选中的项目即出现在表格顶部，如图 3.3 所示。你可以在地图窗口中地图区域外（窗口中的白色部分）任意位置单击取消选择，或从图 3.2 所示的菜单中选择 Clear Selection.

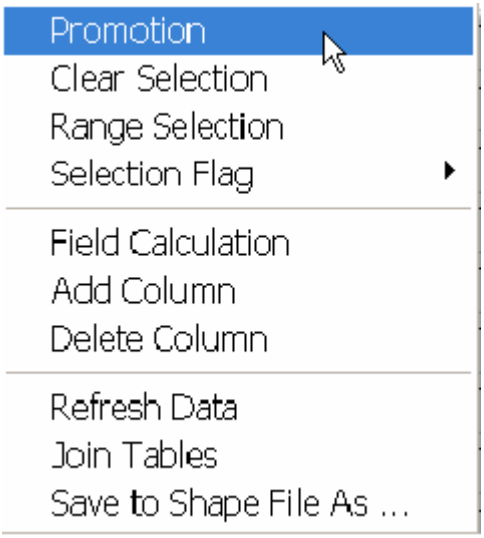


图 3.2　表格下拉菜单

Table : SIDS

	AREA	PERIMETER	CNTY_	CNTY_ID	NAME	STATE_NAME	STATE_FIPS
67	0.181000	1.980000	2040	2040	Moore	North Carolina	37
60	0.065000	1.093000	2026	2026	Lee	North Carolina	37
63	0.154000	1.680000	2030	2030	Harnett	North Carolina	37
54	0.207000	1.851000	1989	1989	Johnston	North Carolina	37
96	0.225000	2.107000	2162	2162	Bladen	North Carolina	37
94	0.240000	2.004000	2150	2150	Robeson	North Carolina	37
82	0.172000	1.835000	2090	2090	Cumberland	North Carolina	37
86	0.098000	1.262000	2097	2097	Hoke	North Carolina	37
79	0.241000	2.214000	2083	2083	Sampson	North Carolina	37
92	0.080000	1.188000	2123	2123	Scotland	North Carolina	37
98	0.240000	2.365000	2232	2232	Columbus	North Carolina	37
90	0.121000	1.855000	2107	2107	Richmond	North Carolina	37

表 3.3　表格中被提升（Promote）的选中行

3.3 表格排序和选择

表格最初的显示方式仅反映了文件中观测对象（observation）的顺序。要根据某一变量的值对观测对象进行排序，在相应变量列的顶部双击。这是一个套索形状：排序顺序在升序排列和降序排列之间转换。一个小的三角形出现在变量名的后面，顶点向上为升序排列，顶点向下为降序排列。以第一列中观测对象编码排序即取消排序。例如，双击 NWBIR74 这一列的顶部，即生成顺序（升序）排列，如图 3.4 所示。

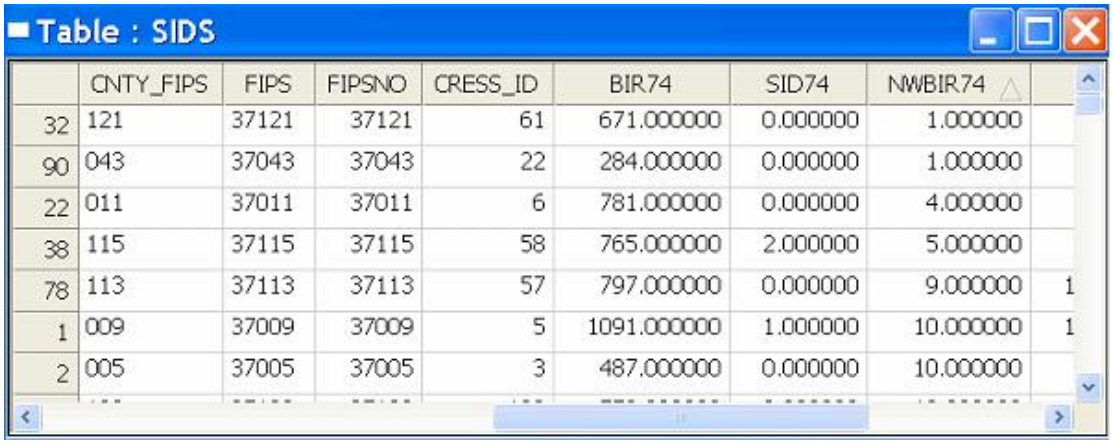


Table : SIDS

	CNTY_FIPS	FIPS	FIPSNO	CRESS_ID	BIR74	SID74	NWBIR74	
32	121	37121	37121	61	671.000000	0.000000	1.000000	
90	043	37043	37043	22	284.000000	0.000000	1.000000	
22	011	37011	37011	6	781.000000	0.000000	4.000000	
38	115	37115	37115	58	765.000000	2.000000	5.000000	
78	113	37113	37113	57	797.000000	0.000000	9.000000	1
1	009	37009	37009	5	1091.000000	1.000000	10.000000	1
2	005	37005	37005	3	487.000000	0.000000	10.000000	

图 3.4　以 NWBIR74 排列的表格

在表格最左边的顺序数字上单击可以选择单独的几行。Shift-click 增加或取消被选对象。你可以在最左边一列拖拽鼠标选择多行记录。被选中的对象立即会在所有链接的地图（及其它图表）中显示出来。右击打开下拉菜单，选择 Clear Selection（或在菜单栏选择 Table>Clear Selection），可以取消选择。

3.3.1 查询

GeaDa 可以进行有限数量的查询，主要适合于选择有特殊值或在某些值之间的对象。可以根据某一变量（只能为一个变量）的变化范围，建立一个逻辑表达式来选择对象。要进行查询，在表格中右击，在下拉菜单中选择 Range Selection（或利用菜单栏中 Table>Range Selection）.出现一个对话窗口要求你输入一个范围（如图 3.5）。注意这一范围包括左侧而不包括右侧（<=和<）。要查找出生人口为 500 或小于 500 的县，在左侧文本框中 0，选择 BIR74　为变量，在右侧文本框中输入 500.1。然后，单击右上部的 Apply 按钮[1]进行查询。

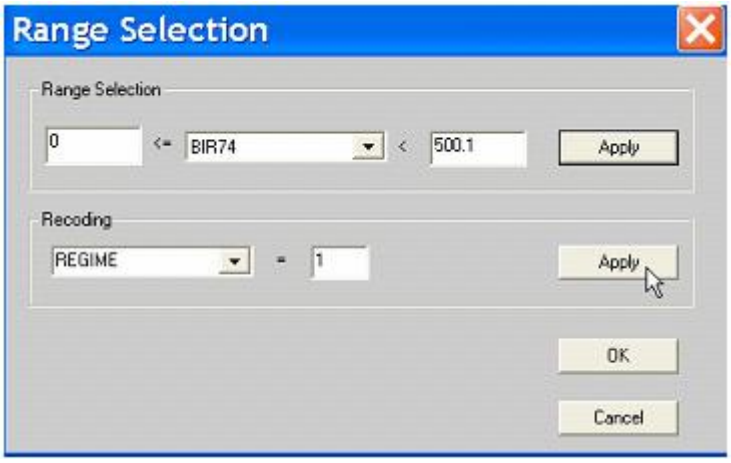


图 3.5　范围选择对话框

第一个 Apply 按钮将会激活 Recoding 对话框，它要求你创建一个新的变量并赋值将被选中的对象为 1，而其它为 0。默认的变量名为 REGIME，但你可以重写文体框来进行修改。如果你不想创建变量，点击 OK。另一方面，如果你希望得到额外的（extra）变量，先单击 Apply（如图 3.5），再单击 OK。

表格中被选中的行将会被用蓝色高亮显示。为把它们放在一起，从下拉菜单中选择 Promotion。结果就会如图 3.6 所示。注意表格中多出（extra）的列变量名为 REGIME。但是这个新变量只是暂时的，只有在保存表之后才会成为永久（参见 3.4）

[1] 图 3.5 中，这个 Apply 按钮在有鼠标的那个 Apply 按钮之上。

Table : SIDS

	FIPS	FIPSNO	CRESS_ID	BIR74	SID74	NWBIR74	BIR79	SID79	NWBIR79	REGIME
87	37095	37095	48	338.000000	0.000000	134.000000	427.000000	0.000000	169.000000	1
2	37005	37005	3	487.000000	0.000000	10.000000	542.000000	3.000000	12.000000	1
7	37029	37029	15	286.000000	0.000000	115.000000	350.000000	2.000000	139.000000	1
73	37075	37075	38	415.000000	0.000000	40.000000	488.000000	1.000000	45.000000	1
20	37143	37143	72	484.000000	1.000000	230.000000	676.000000	0.000000	310.000000	1
90	37043	37043	22	284.000000	0.000000	1.000000	419.000000	0.000000	5.000000	1
8	37073	37073	37	420.000000	0.000000	254.000000	594.000000	2.000000	371.000000	1
45	37177	37177	89	248.000000	0.000000	116.000000	319.000000	0.000000	141.000000	1
9	37185	37185	93	968.000000	4.000000	748.000000	1190.000000	2.000000	844.000000	0

图 3.6　1974 年出生人口少于 500 的县，表视图

3.4 表计算

GeoDa 中，表格中包含了一些有限的计算功能，所以可以添加新的变量，删除现有变量，现有变量的转换等等。你从下拉菜单（右击表格）中选择 Field Calculation 激活（invoke）计算器（见图 3.2）或者从主菜单 Table 中选择 Field Calculation。

计算器顶部有标签，可以选择你希望执行的操作类型。例如图 3.7 中，选中最右边的标签，执行比率计算（operation）。

进行计算之前，你一定（typically）要新建一个新的变量。这需要用 Table 菜单中的 Add Column 命令来激活（或者在表格中右击）。注意这并不是必须的，你可以直接在最左边的文本框 Field Calculation 对话框中输入一个新的变量名（见图 3.7）。这个新字段将增加到表格中。

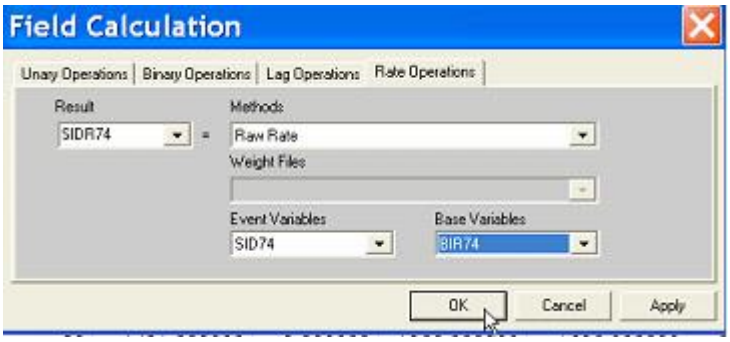


图 3.7　比率计算标签

你可能注意到 sids.shp 文件只包含了出生人口和死亡人口的数字，但没有比值[1]。要新建 1974 年死于 SIDS 的比率，从下拉菜单中选择 Add Column，输入 SIDR74 作为新的变量名，然后点击 Add，如图 3.8 所示。在表格的最右边出现一个新的空列（图 3.9）。

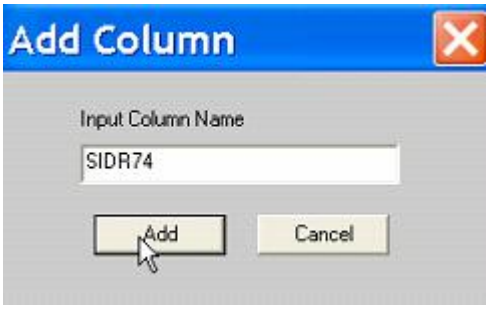


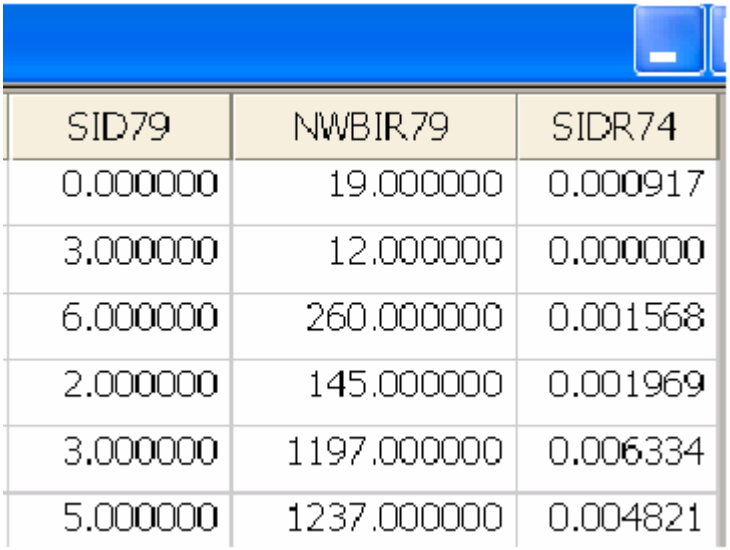
图 3.8　为表格增加一个新的变量

[1] 相反，sids2.shp 样本数据集包括数字及比率。

SID79	NWBIR79	SIDR74
0.000000	19.000000	
3.000000	12.000000	
6.000000	260.000000	
2.000000	145.000000	

图 3.9　带有新的空列的表格

如图 3.10 所示，比率可以并不是最直观的。例如，你可以希望将它调整为人口统计学家或流行病学家所常用的格式，每百万人口的比例。再次激活 Field Calculation，这一次选择第二个标签：Binary Operations。调整变量 SIDR74 为 SIDR74 MUTIPLY 100,000（在变量名 AREA 上输入 100,000），如图 3.11 所示。点击 OK 即完成操作，SIDS 死亡率即被调整后的数值所取代，如图 3.12 所示。



SID79	NWBIR79	SIDR74
0.000000	19.000000	0.000917
3.000000	12.000000	0.000000
6.000000	260.000000	0.001568
2.000000	145.000000	0.001969
3.000000	1197.000000	0.006334
5.000000	1237.000000	0.004821

图 3.10　计算的 SIDS 死亡率被添加到表格中

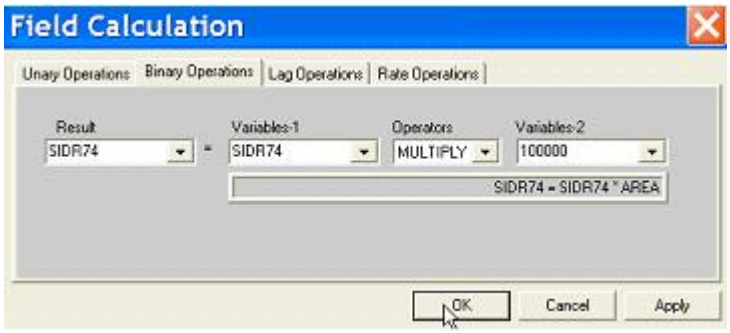


图 3.11　调整 SIDS 人口死亡率

SID79	NWBIR79	SIDR74
0.000000	19.000000	91.700000
3.000000	12.000000	0.000000
6.000000	260.000000	156.800000
2.000000	145.000000	196.900000
3.000000	1197.000000	633.400000

表 3.12　调整后的 SIDS 死亡率被添加到表格中

新计算出的值可以立即被应用于所有地图及统计过程。但一定要记住它们只是暂时的，可以被删除（以防你犯错误）。你通过 Table 菜单或表格的下拉菜单选择 Refresh Data 来完成操作。

只有你将文件以一个不同的名字保存，新的变量才永久保存下来。这个可以通过 Save to Shape File AS 选项来完成[1]。被保存的 shape 文件将使用的地图文件作为当前 shape 文件，但以新近创建的表格作为其 dbf 文件。如果你不需要 shape 文件，你可以稍后删除新的.shp 和.shx 文件，只用 dbf 文件本身（如在电子制表软件或统计软件）。

为 SIDR74 和 SIDR79 创建一个比率变量，将结果保存到一个新的文件中。关闭所有窗口，打开新的 shape 文件查看它的内容。

3.5 练习

关闭所有窗口，载入 St.Louis 的 78 个县的杀人案的样本数据（stl_hom.shp，以 FIPSNO 为关键字）。创建一幅 choropleth 图（如四分位图或标准差图），激活表格。在表格中利用选择工具找出某些县的位置（如在表格中点击 St.Louis 县，查看它在地图中的位置）。对表格进行排序，找出哪些县在 1984 年－1988 年没有发生杀人案（HC8488＝0）。利用范围选择工具找出同时期杀人案小于 5 的县（HC8488＜5）。

为每次选择建立一个虚拟变量（用一个不同的名字，而不是默认的 REGIME）。利用这些新的变量和 Field Calculation 功能（不是 Range Selection），创建另外一次选择，这些县的杀人案大于 0 而小于 5。对不同的（不同时期的）杀人案发生数量（或比率）变量和/或不同的选择范围进行试验。

最后，选择一个时期为 St.Louis 数据建立一个杀人案发生率变量（HCxxxx 和 POxxxx 分别是发生数和基数）。将你计算的结果与表格中已有的（HRxxxx）进行比较。将发生率调整为不同的基数，并将其另存为不同的文件名。关闭所有窗口，载入新的 shape 文件。检查表格并确认所有新的变量都存在。也使用其它的计算选项进行试验。

[1] 这一选项只有在进行计算之后或表格发生变化之后才是可用的。

练习 4　创建一个点文件

4.1 目标

本练习显示你在没有合适的 ESRI 格式的 shape 文件可用，只有 text 或 dbf 文件时，如何创建一个点文件。因为 GeoDa 需要 shape 文件作为输入文件，有时候还是需要这一步的。例如，许多样本数据来自于空间统计表中的文本，这些可以在网上得到，但很少是 shape 文件格式的。不用打开项目即可以完成此项功能（这似乎有些矛盾，因为你没有 shape 文件可以载入）。这可以利用 Tools 菜单来实现。

在本练习后，你将学会：

- 将要输入 GeoDa 的文件进行格式化
- 利用 Text 输入文件或 dbf 数据库文件创建一个点 shape 文件

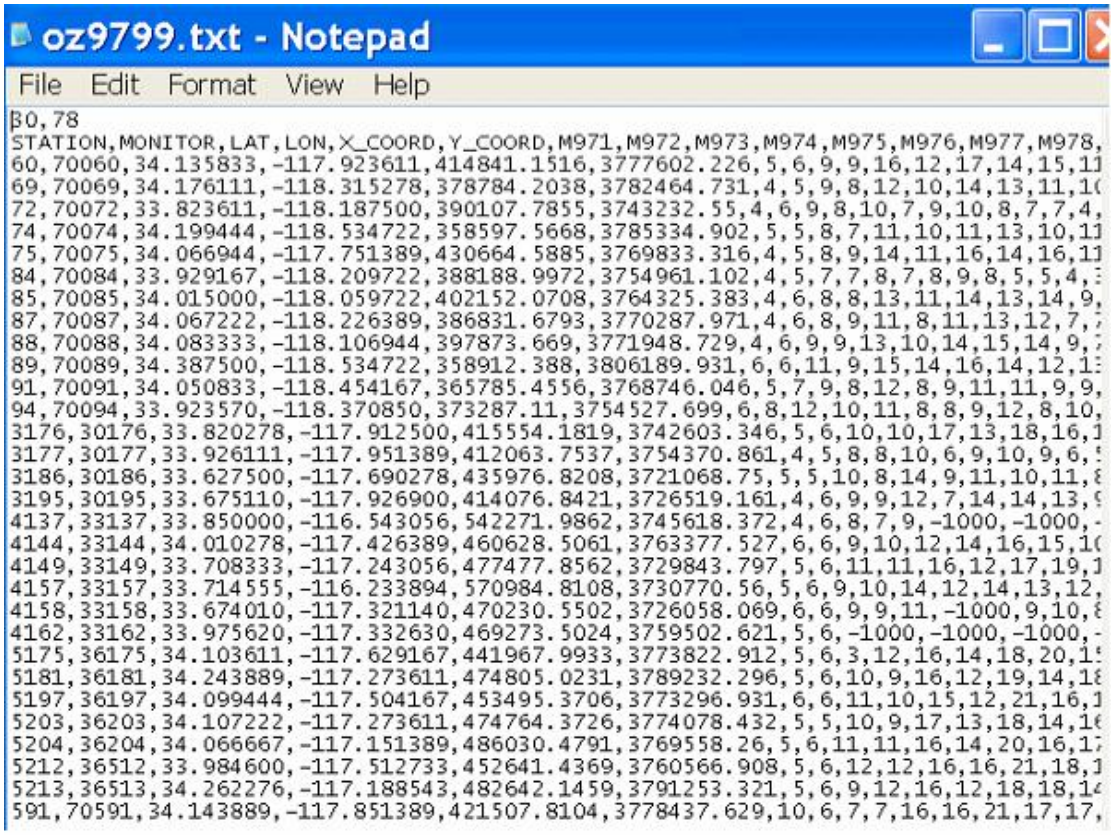
更详细的信息请参阅用户手册第 28－33 页。

4.2 点输入文件格式

用来创建点 shape 文件的输入文件的格式是非常简单的（straightforward）。输入文件最少要包括三个变量：唯一的识别码（整型数值），x 坐标和 y 坐标[1]。在 dbf 格式文件中，没有其它的要求。

当输入文件为 text 格式，每一观察对象的三个必要的变量要单独占一行，用逗号隔开。输入文件必须包括两行头文件。第一行包括观测对象的数目，变量个数，第二行为变量名。两样，所有项目由逗号隔开。除了识别码和坐标，输入文件还包括其它变量[2]。Txet 输入文件的格式如图 4.1 所示，它显示了文本文件 oz9799.txt 中 OZ9799 样本数据的部分内容。该文件包括了 Los Angels 盆地中 30 个监测点采集了每月臭氧（ozone）含量值。第一行给出观测对象的数量（30）和变量的个数（2 个识别码，4 个坐标值和 72 个三年中每月的观测值）。第二行包括了所有的变量名，由逗号隔开。注意包括了未投影的经度和纬度及有投影的 x，y 坐标（UTM zone 11）。

[1] 注意如果有纬度和经度，经度为 x 坐标，纬度为 y 坐标。
[2] 这是相对于练习 5 中用于创建多边形 shape 文件的输入文件而言，需要两步来完成。



oz9799.txt - Notepad

File Edit Format View Help

```
β0,78
STATION,MONITOR,LAT,LON,X_COORD,Y_COORD,M971,M972,M973,M974,M975,M976,M977,M978,
60,70060,34.135833,-117.923611,414841.1516,3777602.226,5,6,9,9,16,12,17,14,15,11
69,70069,34.176111,-118.315278,378784.2038,3782464.731,4,5,9,8,12,10,14,13,11,1(
72,70072,33.823611,-118.187500,390107.7855,3743232.55,4,6,9,8,10,7,9,10,8,7,7,4,
74,70074,34.199444,-118.534722,358597.5668,3785334.902,5,5,8,7,11,10,11,13,10,11
75,70075,34.066944,-117.751389,430664.5885,3769833.316,4,5,8,9,14,11,16,14,16,11
84,70084,33.929167,-118.209722,388188.9972,3754961.102,4,5,7,7,8,7,8,9,8,5,5,4,
85,70085,34.015000,-118.059722,402152.0708,3764325.383,4,6,8,8,13,11,14,13,14,9,
87,70087,34.067222,-118.226389,386831.6793,3770287.971,4,6,8,9,11,8,11,13,12,7,
88,70088,34.083333,-118.106944,397873.669,3771948.729,4,6,9,9,13,10,14,15,14,9,
89,70089,34.387500,-118.534722,358912.388,3806189.931,6,6,11,9,15,14,16,14,12,1
91,70091,34.050833,-118.454167,365785.4556,3768746.046,5,7,9,8,12,8,9,11,11,9,9,
94,70094,33.923570,-118.370850,373287.11,3754527.699,6,8,12,10,11,8,8,9,12,8,10,
3176,30176,33.820278,-117.912500,415554.1819,3742603.346,5,6,10,10,17,13,18,16,1
3177,30177,33.926111,-117.951389,412063.7537,3754370.861,4,5,8,8,10,6,9,10,9,6,
3186,30186,33.627500,-117.690278,435976.8208,3721068.75,5,5,10,8,14,9,11,10,11,
3195,30195,33.675110,-117.926900,414076.8421,3726519.161,4,6,9,9,12,7,14,14,13,
4137,33137,33.850000,-116.543056,542271.9862,3745618.372,4,6,8,7,9,-1000,-1000,-
4144,33144,34.010278,-117.426389,460628.5061,3763377.527,6,6,9,10,12,14,16,15,1(
4149,33149,33.708333,-117.243056,477477.8562,3729843.797,5,6,11,11,16,12,17,19,1
4157,33157,33.714555,-116.233894,570984.8108,3730770.56,5,6,9,10,14,12,14,13,12,
4158,33158,33.674010,-117.321140,470230.5502,3726058.069,6,6,9,9,11,-1000,9,10,
4162,33162,33.975620,-117.332630,469273.5024,3759502.621,5,6,-1000,-1000,-1000,-
5175,36175,34.103611,-117.629167,441967.9933,3773822.912,5,6,3,12,16,14,18,20,1
5181,36181,34.243889,-117.273611,474805.0231,3789232.296,5,6,10,9,16,12,19,14,1
5197,36197,34.099444,-117.504167,453495.3706,3773296.931,6,6,11,10,15,12,21,16,1
5203,36203,34.107222,-117.273611,474764.3726,3774078.432,5,5,10,9,17,13,18,14,1
5204,36204,34.066667,-117.151389,486030.4791,3769558.26,5,6,11,11,16,14,20,16,1
5212,36512,33.984600,-117.512733,452641.4369,3760566.908,5,6,12,12,16,16,21,18,1
5213,36513,34.262276,-117.188543,482642.1459,3791253.321,5,6,9,12,16,12,18,18,1
591,70591,34.143889,-117.851389,421507.8104,3778437.629,10,6,7,7,16,16,21,17,17,
```

图 4.1　具有坐标的 Los Angeles 臭氧数据文本的文件

4.3 将 text 文件转换为一个点 shape 文件

由 text 文件创建点 shape 文件要用 Tools 菜单，选择 Shape>Points from ASCII，如图 4.2 所示。当输入文件为 dbf 文件，相应的命令为 Shape>Points from DBF。这将产生一个对话框，要求指定一个 text 文件路径及新的 shape 文件名。前面输入 oz9799.txt，后面输入 oz9799（文件名后缀 shp 将由程序自行添加）。然后，要为 oz9799.txt 文件中具有投影的坐标设置 X 坐标和 Y 坐标，如图 4.3 所示。利用这些相同的值或选择 LON 和 LAT。点击 Create 按钮将生成一个 shape 文件。最后点击 OK 将返回主界面。

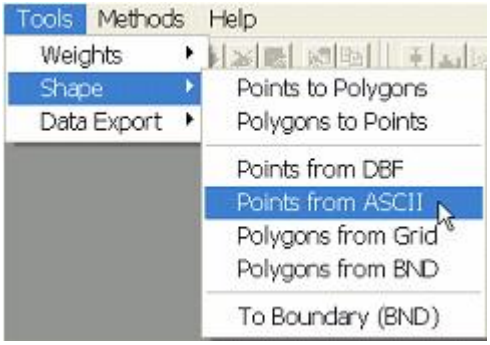


图 4.2　由 text 文件创建一个点 shape 文件

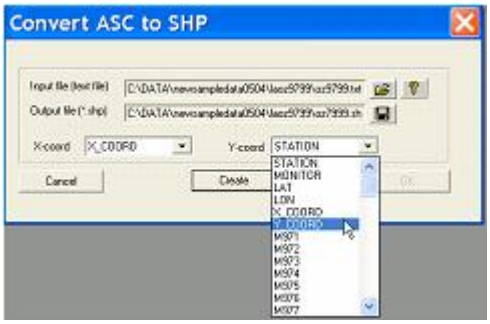
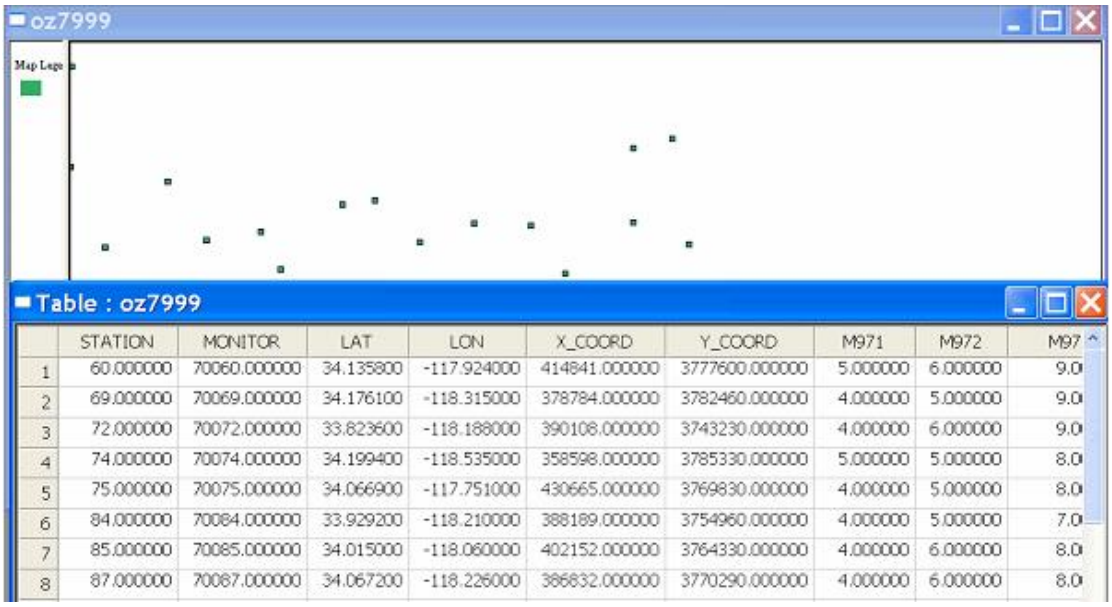


图 4.3　为点 shape 文件选择 x 坐标和 y 坐标

打开一个新项目（File>Open Project），选择 oz9799.shp 文件，检查新创建的 shape 文件的内容。点地图及相连的数据表格如图 4.4 所示。注意，相对于 ESRI 点文件标准，点的坐标是包含在数据表格中的。



	STATION	MONITOR	LAT	LON	X_COORD	Y_COORD	M971	M972	M97
1	60.000000	70060.000000	34.135800	-117.924000	414841.000000	3777600.000000	5.000000	6.000000	9.0
2	69.000000	70069.000000	34.176100	-118.315000	378784.000000	3782460.000000	4.000000	5.000000	9.0
3	72.000000	70072.000000	33.823600	-118.188000	390108.000000	3743230.000000	4.000000	6.000000	9.0
4	74.000000	70074.000000	34.199400	-118.535000	358598.000000	3785330.000000	5.000000	5.000000	8.0
5	75.000000	70075.000000	34.066900	-117.751000	430665.000000	3769830.000000	4.000000	5.000000	8.0
6	84.000000	70084.000000	33.929200	-118.210000	388189.000000	3754960.000000	4.000000	5.000000	7.0
7	85.000000	70085.000000	34.015000	-118.060000	402152.000000	3764330.000000	4.000000	6.000000	8.0
8	87.000000	70087.000000	34.067200	-118.226000	386832.000000	3770290.000000	4.000000	6.000000	8.0

图 4.4　OZ9799 点 shape 文件底图和数据表

4.4 练习

文件 BOSTON 包括典型 Harrison 和 Rubinfeld（1978）的房屋数据，包括 506 个人口区的 23 个变量。原始数据利用地区中心的坐标进行了放大，包括未投影的经度和纬度和有投影的 x，y 坐标。利用 boston.txt 文件创建一个房屋数据的点 shape 文件。你也可以实验用 dbf 文件创建点 shape 文件，如 BALTIMORE，JUVENILE 和 PITTSBURGH。

练习 5　创建一个多边形文件

5.1 目标

本练习将会向显示在你没有合适的 ESRI 格式文件时，你如何利用文本文件来创建一个不规则格网或规则格网的多边形 shape 文件。如练习 4 中所示，该功能可以通过 Tools 菜单来实现，而不必打开一个项目。

在本练习之后，你会知道：

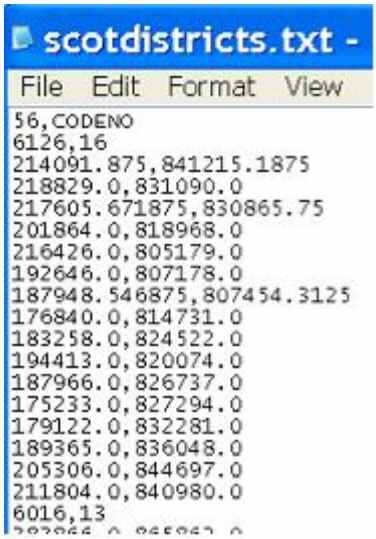
- 如何利用带有边界坐标的文件创建一个多边形文件。
- 创建一个规则格网的多边形文件
- 将数据表链接到底图文件中

有关本操作的更详细信息请参阅发行注解（Release Notes）13—17 页或用户手册第 63—64 页。

5.2 边界文件输入格式

目前 GeoDa 支持一种多边形边界坐标输入文件格式。虽然这是一种限制，但实际中这一文件可以直接从一种格式转换为其它格式。它所支持的格式，如图 5.1 所示，由包含多边形数量的头行和唯一的多边形识别码组成，这些识别码由逗号分开。对于每一个多边形，都列出了它的识别码和点的数量，以及每个点的坐标对（由逗号分开）。这种格式 User Guide 中被称为 1a 格式。注意它目前还不支持带有相同观测的多边形。同样的，第一个点的座标对不能在最后重复。每个多边形的点坐标的数量反映了这一点（图 5.1 中第一个多边形有 16 对 X，Y 坐标）。

图 5.1 中的边界文件是关于典型 Scottish 唇癌的数据，这些数据在许多文章中都有（参见 Cressie 1993，第 573 页）。这 56 个行政区的坐标来自于 WinBugs 软件包中 scotland.map 边界，被导出为 S-Plus 地图格式。然后将结果文件编辑为与 GeoDa 输入格式相一致的格式。另外，重复的坐标和碎多边形被清除。将结果保存于 scotdistricts.txt 文件中。注意避免多边形带来的问题，岛状多边形被简化为单多边形。


scotdistricts.txt -

File Edit Format View

```
56,CODENO
6126,16
214091.875,841215.1875
218829.0,831090.0
217605.671875,830865.75
201864.0,818968.0
216426.0,805179.0
192646.0,807178.0
187948.546875,807454.3125
176840.0,814731.0
183258.0,824522.0
194413.0,820074.0
187966.0,826737.0
175233.0,827294.0
179122.0,832281.0
189365.0,836048.0
205306.0,844697.0
211804.0,840980.0
6016,13
```

图 5.1 带有 Scottish 地区边界座标的输入文件

相对于练习 4 中对点 shape 文件的程序，这里还需要两步。第一，建立一幅底图（见 5.3）。这个文件除了多边形识别码，面积和周长外不包含其它任何数据。第二步，将数据表链接到这个 shape 文件中，添加感兴趣的变量（见 5.4）。

5.3 为底图创建一个多边形文件

在 Tools 菜单中选择 Shape>Polygons from BND 来创建底图，如图 5.2 所示。出现一个对话框，如图 5.3 所示，指定输入文件的路径，及新建 shape 文件的文件名。前者选择 scotdistricts.txt，输入 scotdistricts 作为底图 shape 文件的文件名。然后，点击 Create 启动创建程序。当蓝色进度条（见图 5.3）显示转换完成，点击 OK 返回到主菜单。

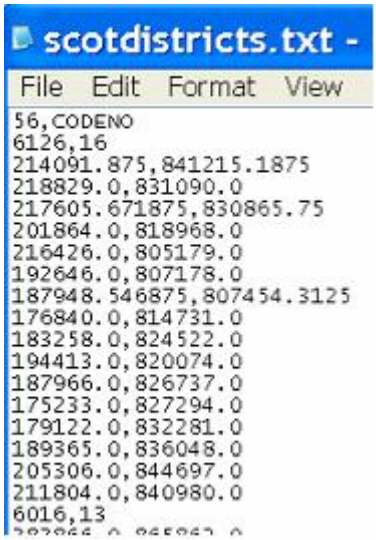


图 5.2 由 ascii 文本文件创建一个多边形 shape 文件

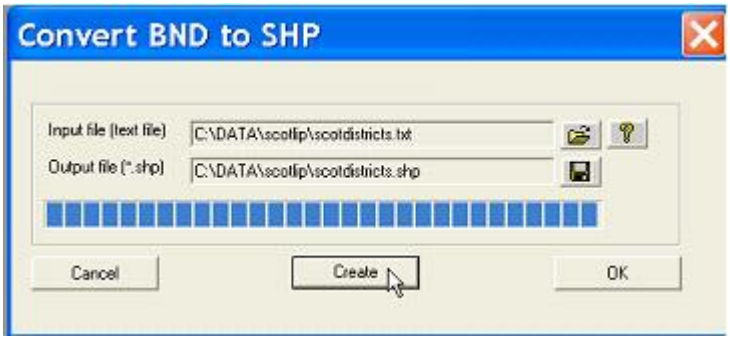


图 5.3 指定 Scottish 地区输入与输出文件

结果底图如图 5.4 所示，此图是通过 Open project 工具按钮创建的，之后输入文件名，将 CODENO 作为主变量。然后点击 Table 工具条按钮来打开相应的数据表。如图 5.5 所示，这只包括识别码和一些几何信息，但没有其它有用的数据。

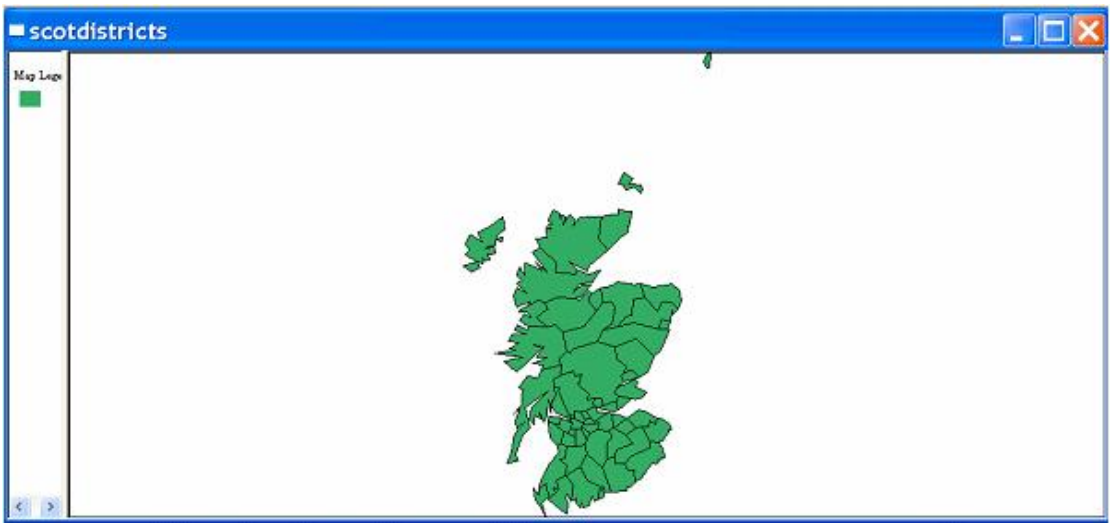


图 5.4 Scottish 地区底图

Table : scotdistricts

	CODENO	AREA	PERIMETER	RECORD_ID
1	6126	974002000.000000	184951.000000	1
2	6016	1461990000.000000	178224.000000	2
3	6121	1753090000.000000	179177.000000	3
4	5601	898599000.000000	128777.000000	4
5	6125	5109870000.000000	580792.000000	5
6	6554	422639000.000000	118433.000000	6
7	6019	2267340000.000000	259143.000000	7
8	6655	157575000.000000	54859.000000	8
9	6123	4129180000.000000	543208.000000	9
10	6017	2084800000.000000	260668.000000	10
11	6756	2118860000.000000	357924.000000	11
12	6127	6031740000.000000	458473.000000	12
13	6124	392932000.000000	80920.500000	13
14	5811	1493140000.000000	309530.000000	14
15	5914	757753000.000000	122569.000000	15

图 5.5 Scottish 地区底图数据表

5.4 为底图连接一个数据表

为创建 Scottish 地区包含唇癌数据的 shape 文件，必须为底图连接一个数据表（dbf 格式）。这可以使用 Table 菜单中 Join Tables 命令（或在表格中右击，从下拉菜单中选择 Join Tables，如图 3.2 所示）。

出现一个 Join Tables 对话框，如图 5.6 所示。输入所导入的文件名为 scotlipdate.dbf，选择 CODENO 为主变量，如图所示。然后通过点击>>按钮将左边一栏的所示变量移动到右边，如图 5.7 所示。最后，点击 Join 按钮完成操作。结果数据表如图 5.8 所示。

图 5.6 指定连接的数据表及主变量

图 5.7 选择连接表中的变量

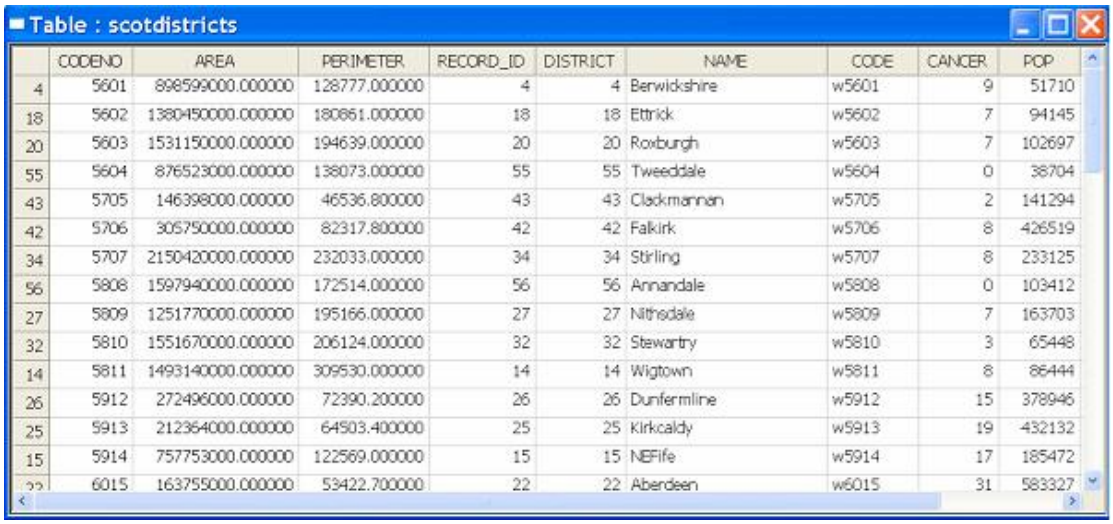


Table : scotdistricts

	CODENO	AREA	PERIMETER	RECORD_ID	DISTRICT	NAME	CODE	CANCER	POP
4	5601	898599000.000000	128777.000000	4	4	Berwickshire	w5601	9	51710
18	5602	1380450000.000000	180861.000000	18	18	Ettrick	w5602	7	94145
20	5603	1531150000.000000	194639.000000	20	20	Roxburgh	w5603	7	102697
55	5604	876523000.000000	138073.000000	55	55	Tweeddale	w5604	0	38704
43	5705	146398000.000000	46536.800000	43	43	Clackmannan	w5705	2	141294
42	5706	305750000.000000	82317.800000	42	42	Falkirk	w5706	8	426519
34	5707	2150420000.000000	232033.000000	34	34	Stirling	w5707	8	233125
56	5808	1597940000.000000	172514.000000	56	56	Annandale	w5808	0	103412
27	5809	1251770000.000000	195166.000000	27	27	Nithsdale	w5809	7	163703
32	5810	1551670000.000000	206124.000000	32	32	Stewartry	w5810	3	65448
14	5811	1493140000.000000	309530.000000	14	14	Wigtown	w5811	8	86444
26	5912	272496000.000000	72390.200000	26	26	Dunfermline	w5912	15	378946
25	5913	212364000.000000	64503.400000	25	25	Kirkcaldy	w5913	19	432132
15	5914	757753000.000000	122569.000000	15	15	NEFife	w5914	17	185472
22	6015	163755000.000000	53422.700000	22	22	Aberdeen	w6015	31	583327

图 5.8 将 Scottish 唇癌数据库连接到底图中

图 5.8 所示的表格中所有变量都是可以用来制图及分析。但是要使这些变量永久保持下来，必须用一个新的文件名保存文件，如 3.4 所述。可以用 Table 菜单中 Save to Shape File As..或在表格中右击来实现保存文件，如图 5.9。选择这一命令，为输出的 shape 文件输入一个新的文件名（如 scotdistricts），然后点击 OK。关闭项目，载入这一新的 shape 文件，查看它的内容如预料的一样。

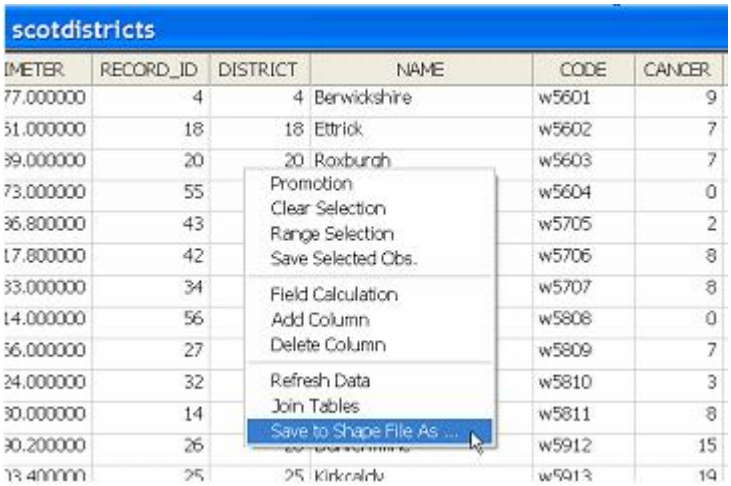


图 5.9　保存连接的 Scottish 唇癌数据为一个新的 shape 文件

5.5 创建一个规则格网多边形 shape 文件

不必指定边界的真实坐标，GeoDa 可以创建规则格网（或网格）的多边形 shape 文件。从 Tools 菜单，利用 Shape>Polygons from Grid 功能来实现，如图 5.10 所示。

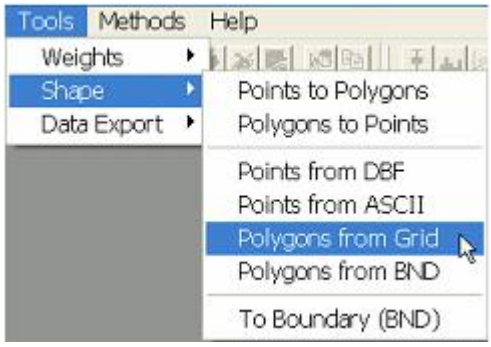


图 5.10 创建一个多边形格网的 shape 文件

这会激活一个对话框，提供许多指定格网显示的不同属性，如图 5.11 所示。我们这里只谈最简单的（更详细的内容参见 Release Notes）。

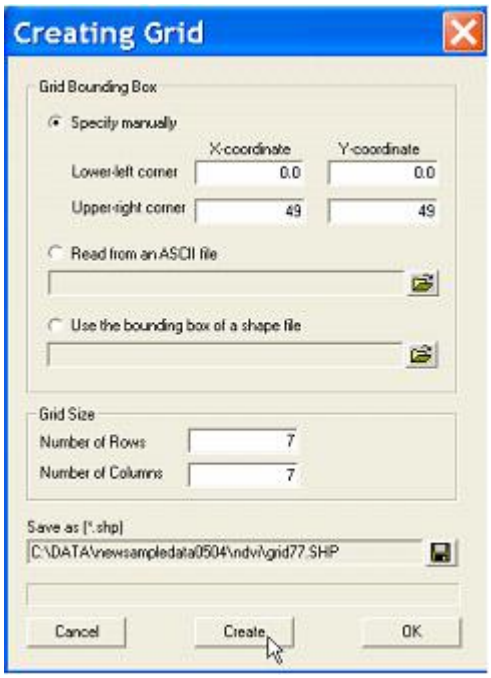


图 5.11　指定规则格网的维数

如图 5.11 所示，点击单选按钮，然后手工输入左下角的坐标为默认设置 0.0，0.0，右上角的坐标为 49,49。在 Grid Size 文本框中，输入行和列数为 7。最后确定 shape 文件的文件名，如 grid77（见图 5.11）。点击 Create 按钮，然后点击 OK，返回至主菜单。

通过 Open project 工具栏按钮，以 PolyID 为主变量，查看格网结果文件。其形状会如图 5.12 所示。利用 Table 工具按钮打开关联数据表。注意它只包括 POLYID 识别码和两个几何特征，如图 5.13 所示。

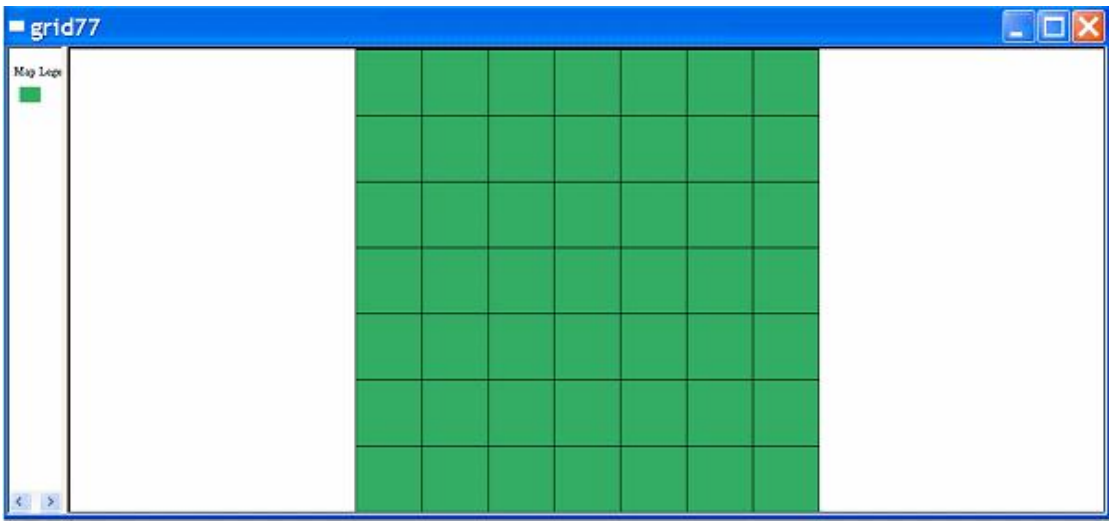


图 5.12　7×7 的规则正方形底图

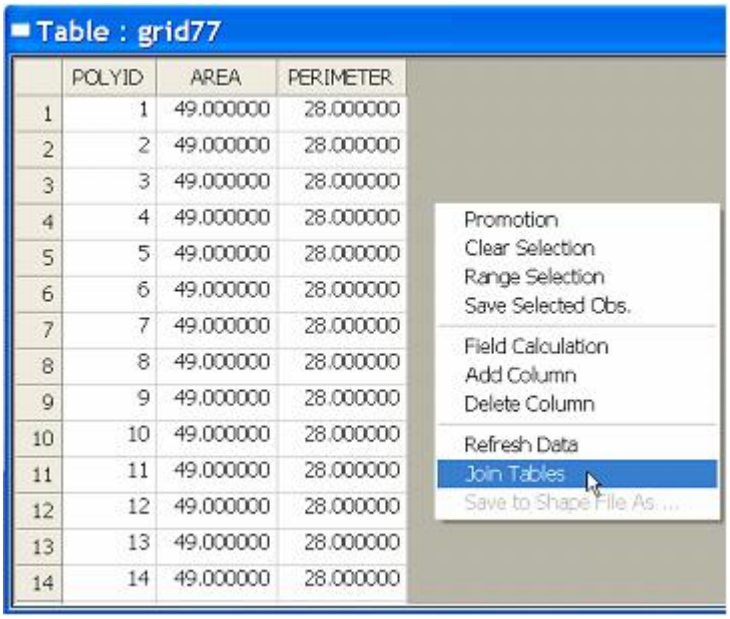


图 5.13　将 NDVI 数据表连接到格网底图

如 5.4 部分，你将需要为这一表格连接一个真实的数据表，以得到一个有意义的项目。选择 Join Table 功能，指定 ndvi.dbf 文件作为输入文件。这个文件从全球变化数据库中提取的，包括为 7×7 正方形栅格而测定的 4 个变量，间隔为 10 弧分。这个文件被 Anselin(1993)所演示。49 个观测点与刚刚建立的规则格网相匹配。

除了文件名，选择 POLYID 作为主变量，将 4 个变量移到右边的栏中，如图 5.14 所示。最后点击 Join 按钮来执行连接（join）。新的数据表包括 4 个新变量，如图 5.15。输入一个新文件名如 ndvigrid，保存文件来完成此程序。

关闭文件之后，打开新文件查看其内容。

图 5.14　指定要连接的 NDVI 变量

Table : grid77

	POLYID	AREA	PERIMETER	GREEN	TEMP	ELEV	PREC
1	1	49.000000	28.000000	126	277	730	107
2	2	49.000000	28.000000	130	274	670	106
3	3	49.000000	28.000000	129	271	610	104
4	4	49.000000	28.000000	126	271	610	104
5	5	49.000000	28.000000	125	276	610	89
6	6	49.000000	28.000000	125	281	610	74
7	7	49.000000	28.000000	122	281	670	74
8	8	49.000000	28.000000	132	271	730	114
9	9	49.000000	28.000000	130	268	670	115

图 5.15 连接规则格网底图了的 NDVI 数据库

5.6 练习

样本数据集中包括许多文件，可以用来练习本章所包括的操作。OHIOLUNG 数据集包括 17 区的 Ohio 州的 88 个县、投影为 UTM 的文本文件 ohiooutmbnd.txt。利用此文件创建一个多边形文件。然后将文件 ohdat.dbf 中“典型”Ohio 肺癌死亡数据（Xia and Carlin，1998）加入此文件。利用 FIPSNO 作为主变量，创建一个包括所有变量的 shape 文件[1]。

或者，你可以应用 Tools>Shape>To Boundare(BND)功能，为每一个多边形 shape 文件创建 1a 格式的坐标文本文件。这可以被用来为该文件再创建与 dbf 文件相关联的多边形文件。

GRID100 样本数据集包括文件 grid10×10.dbf，该文件包括规则 10×10 正方形格网模拟空间自相关随机变量。创建这样一个格网，连接数据文件（主变量为 POLYID）。将结果保存为一个新的文件，你可以用来为进行空间自回归或空间移动平均处理的变量不同模式制图[2]。

[1] Ohlung.shp/shx/dbf 文件中包含结果。
[2] Grid100s.shp/shx/dbf 文件中包含结果。

或者，试验创建格网数据集与样本数据集中 bounding box 相匹配。如，利用 COLUMBUS 图创建一个 7×7 的 Columbus 数据，或利用 SIDS 图创建一个 5×20 的 North Carolina 数据。试验图 5.11 的对话框中提供的不同选项。

练习 6　空间数据操作

6.1 目标

本练习将显示你如何通过计算多边形质心和将 Thiessen 多边形应用于点，在点与多边形间改变空间对象表现形式[1]。如练习 4 和练习 5，可以不打开项目，通过 Tools 菜单实现这一功能。注意这些操作后台的计算只对投影坐标有效，因为它们是在 Euclidean 平面上运行的。因为它们是在 lat-lon 坐标下工作的（GeoDa 不能显示出坐标是否有投影），结果只能是大概，将不能用于精确分析。

在本练习后，你将会知道：

- 创建一个包含质心的点文件
- 将多边形质心加入到当前数据表中
- 创建包含 Thiessen 多边形的多边形文件

有关本操作的更详细的信息见 Users Guide 第 19-28 页，和 Release Notes 第 20-21 页。

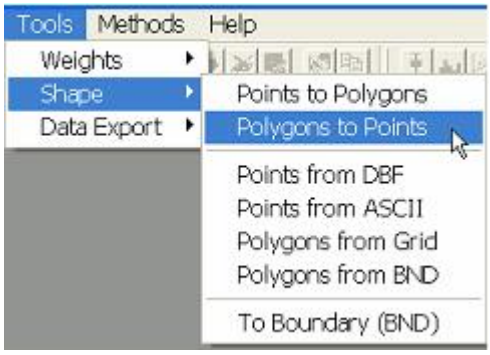


图 6.1 创建一个包含多边形的点文件

图 6.2 指定多边形输入文件

图 6.3 指定点输出文件

[1] 更准确地说，GeoDa 中所提到的质心是中心点，或多边形边界 X 和 Y 坐标的平均数。

6.2 创建一个包含质心坐标的点文件

质心坐标可以在不打开项目的前提下被转为一个点文件。从 Tools 菜单中，选择 Shape>Polygons to Points(图 6.1)来打开 Shape Conversion 对话框。首选，指定多边形输入文件的文件名，如 ohlung.shp，如图 6.2 所示(点击文件打开按钮，打开熟悉的文件对话框)。一旦输入文件名，一幅小的 88 个 Ohio 州的概要图即出现在对话框左边的框中(图 6.2)。然后，输入新文件的文件名，如图 6.3 中 ohcent，点击 Creat 按钮。

新文件创建之后，它的略图会出现在对话框右边的框中如图 6.4 所示。点击 Done 按钮返回主界面。

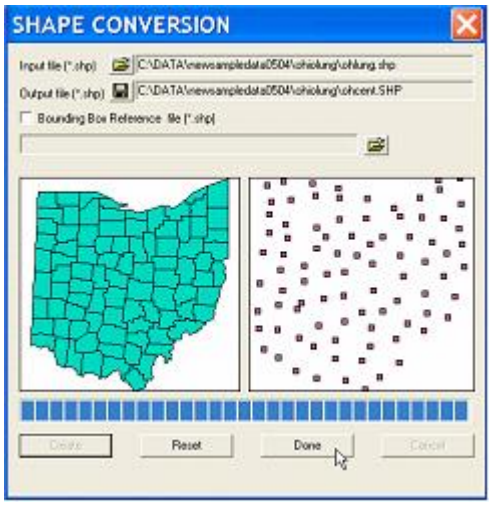


图 6.4 创建质心文件

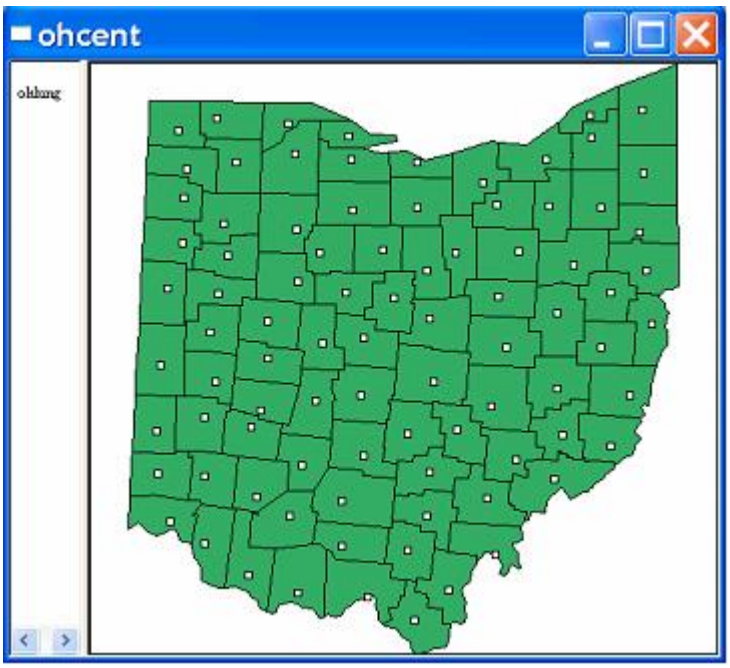


图 6.5 质心点文件覆盖于原始 Ohio 州之上

查看新的文件，首选打开原始 Ohio 州的项目(ohlung.shp，以 FIPSNO 为主变量)。改变 Map 颜色为白色(见图 1.4 的对话框)。然后添加一个带有质心文件(ohcent，以 FIPSNO 为关键字)的新的图层(点击 Add a layer 工具条按钮，或利用菜单中的 Edit>Add Layer)。原始多边形会被质心覆盖，如图 6.5。多边形的白色地图背景被转换成为“点”。作为最上一层，它具有地图所具有的所有属性。

查看数据表的内容。它和原始文件是完全相同的，只是质心坐标作为一个变量加入其中。

6.2.1 在数据表中添加质心坐标

不必创建一个新文件，就可以将多边形质心坐标加入到多边形文件的数据表中。当你想利用这些坐标进行统计分析时，这一点是很有用的(如在趋势面回归中，见 23.3 节)。

该功能是作为地图选项之一，可以从 Option 菜单调用(以一幅地图作为活动窗口)，或在地图中右击，选择 Add Centroids to Table，如图 6.6 所示。或者，有一工具按钮来实现同样的功能。

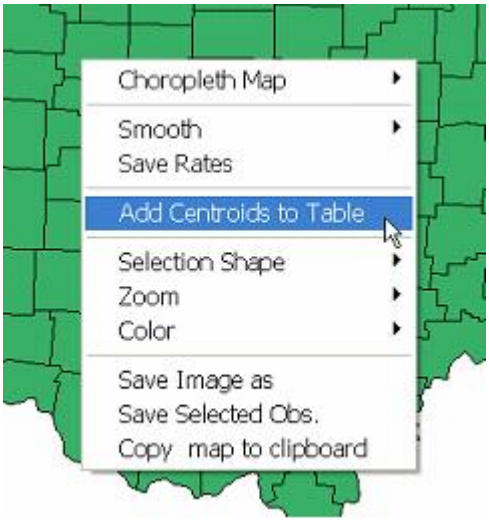


图 6.6 从当前多边形文件向数据表中添加质心

载入(或重新载入)Ohio Lung cancer 数据集(ohlung.shp，以 FIPSNO 为关键字)，选择 Add Centroids to Table 选项。打开一个对话框来为 X 和 Y 坐标指定变量名，如图 6.7 所示。注意你不必为两个坐标都指定变量名，另一坐标也会被选中。保持默认值为 XC00 和 YC00，点击 OK，添加新变量。新数据表会显示，如图 6.8 所示。如前所示，为了使新增变量成为永久变量，要保存为一个新文件。

图 6.7 为质心坐标指定变量名

LF88	POPF88	XCOO	YCOO
116	241573	278258.692308	4607537.692308
2	19509	230392.125000	4610927.500000
12	39013	482102.642857	4597812.142857
7	18616	204053.555556	4603143.333333
352	760016	451944.631579	4583588.947368

图 6.8 添加到数据表中的 Ohio 质心坐标

6.3 创建一个 Thiessen 多边形文件

点文件可以通过 Thiessen 多边形格网(tessellateion)的方法转换为多边形。多边形显示通常

用于变量的空间分布的可视化，也可以建立基于邻接关系的空间权重。这一过程可以通过 Tools 菜单选择 Shape>Points to Polygons 来实现，如图 6.9。

这将打开一个对话框，如图 6.10 所示。指定输入(点)文件 oz9799.shp，这是 30 个洛杉矶流域空气质量监测点位置的数据集。如多边形到点的转换，指定输入文件名会在对话框左边面板中出现一个缩略图。然后输入新(多边形)文件名，为 ozthies.shp。

点击 Create，看到一个 Thiessen 多边形缩略图出现在右边的面板中(图 6.11)。最后选择 Done 返回标准界面。

将 Thiessen 多边形与原始点模式进行比较，如 6.2 节中质心。首先打开 Thiessen 多边形文件(ozthies，以 Station 为关键字)。改变地图颜色为白色。然后添加原始点图层(oz9799，以 Station 为关键字)。结果如图 6.12。查看数据表内容。它与点分布是完全相同的，多了 Area 和 Perimeter 两个变量。

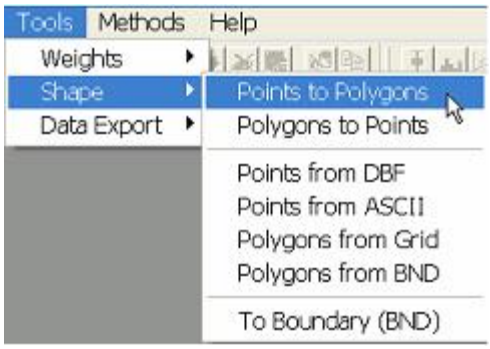


图 6.9 利用点文件创建一个 Thiessen 多边形文件

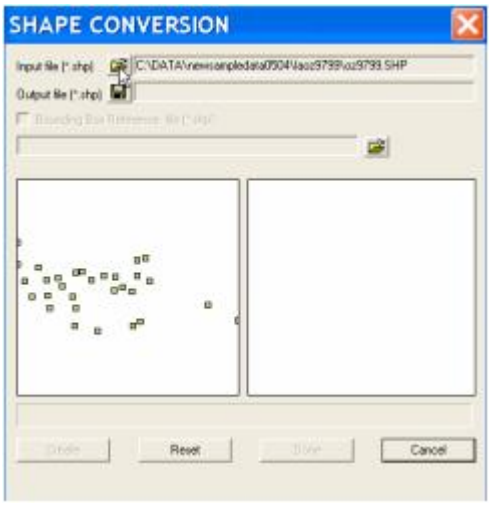


图 6.10 指定点输入文件

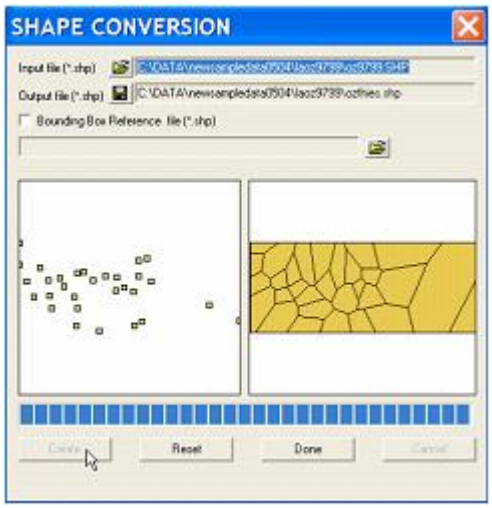


图 6.11 指定 Thiessen 多边形输出文件

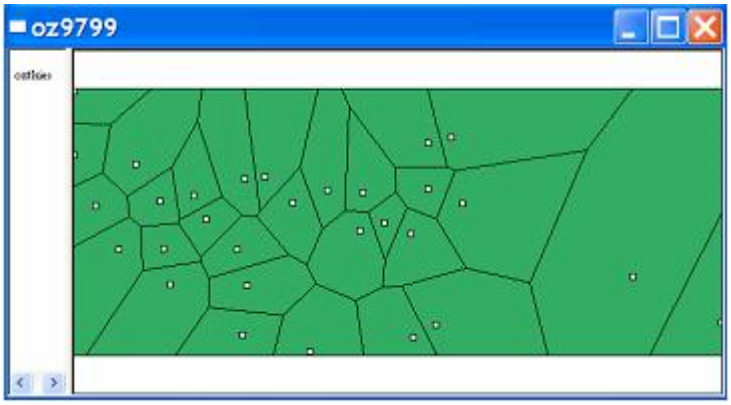


图 6.12 洛杉矶流域监测点的 Thiessen 多边形

6.4 练习

利用 SCOTLIP 数据集建立一个 56 个 Scottish 地区质心的点文件。用这些点产生一个 Thiessen 多边形文件，与原始显示相比较。你也可以用其它样本数据作实验(但是要记住，质心和 Thiessen 多边形文件的结果对无投影和 Lat-lon 坐标是不可靠的)。

或者，打开一个点文件，如 BOSTON 数据集(关键字是 ID)中 506 个人口区的质心，或 BALTIMORE 数据集(关键字是 STATION)中 211 个住房位置。这些都是有投影的坐标。将它们转换为多边形区域。用多边形为各自的房屋中值(MEDV)或房屋价格(PRICE)创建一个面量(分区统计 ，等值区域线图，choropleth)图。用原始点文件与 choropleth 比较一下。

练习 7　EDA 基础，链接

7.1 目标

该练习显示了探索数据分析(EDA)的一些基本方法。包括通过柱状图、箱图对数据非空间分布进行可视化，它强调链接的概念，链接是 GeoDa 的基础。

在本练习后，你会掌握：

- 为变量创建一个柱状图
- 改变柱状图中分类数量
- 创建一个区域柱状图
- 为变量创建一个箱图
- 在箱图中改变标准来决定离异点
- 在柱状图、箱图和地图之间建立链接

关于这些操作更详细的内容请参见 Users Guide，和 Release Notes。

7.2 链接柱状图

我们在柱状图例中，变量非空间分布的可视化来开始传统 EDA 的演示。柱状图是随机变量密度函数的离散近似值，用于探测不对称性，多模态(multiple modes)和其它分布特征。关闭所有窗口，打开一个新的项目，GRID100S 数据集(输入 grid100s，关键字为 PolyID)。开始创建两个五分位图(Map>Quantile，分类数为 5；详见练习 2)，一个是 zar09，一个是 ranzar09[1]。结果应该如图 7.1。注意在左边高度空间自相关相联系的特征丛聚，相反的右边似乎为随机分布模式。

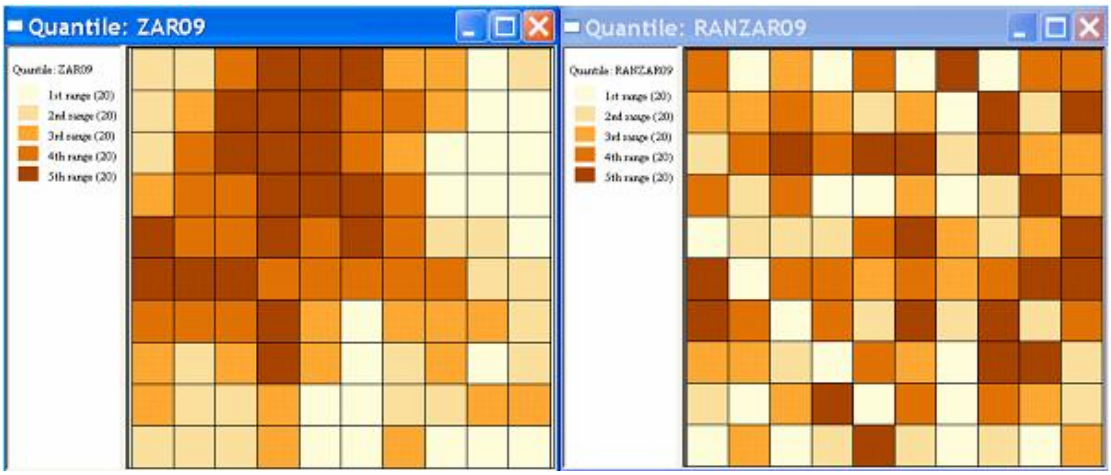


图 7.1 空间 AR 变量的 10×10 五分位图

在菜单(如图 7.2)中 Explore>Histogram 调用柱状图函数，或点击 Histogram 工具按钮。在变量设置对话框，选择 zar09，如图 7.3。结果会出现一个分为 7 类的柱状图，如图 7.4。这显示出正态随机分布变量的钟形形状，随值的变量出现边疆的颜色渐变。柱条顶部的数据显示落在每一间隔

[1] 第一个变量 zar09 描述了 10×10 正方形格网空间自相关过程，参数为 0.9，ranzar09 是相同数值的随机序列。

的观察点数目。间隔本身显示在右侧。

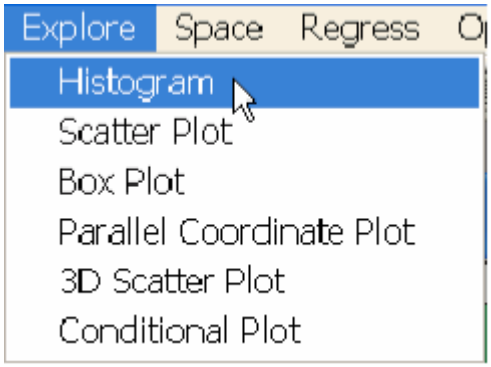


图 7.2 柱状图函数

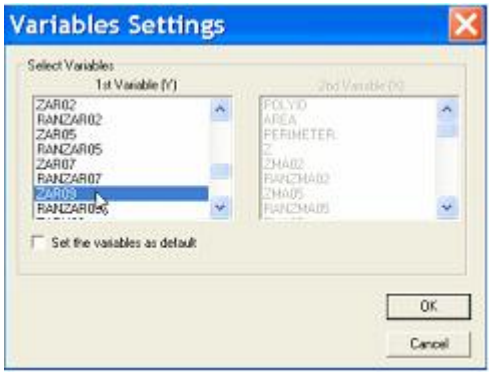


图 7.3 为柱状图选择变量

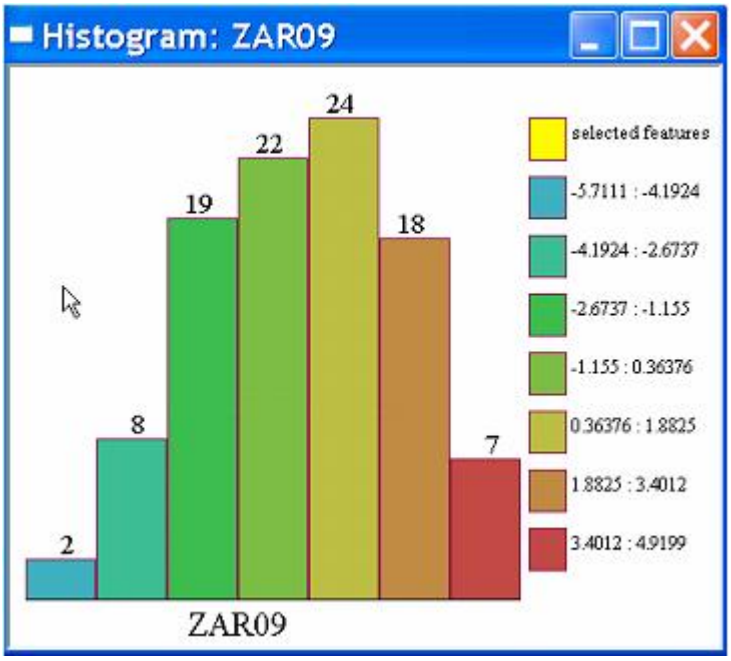


图 7.4 随机变量空间自相关柱状图

现在，利用变量 ranzar09 重复这一过程。比较图 7.5 中这两个柱状图。即使在图 7.1 中表现出不同的空间模式，两个变量的柱状图是相同的。你可以通过比较每一类别的观察点数目和类别数值范围来验证。换句话说，这两个变量的唯一区别即是变量值位于什么地方，而不是分布无空间特征。

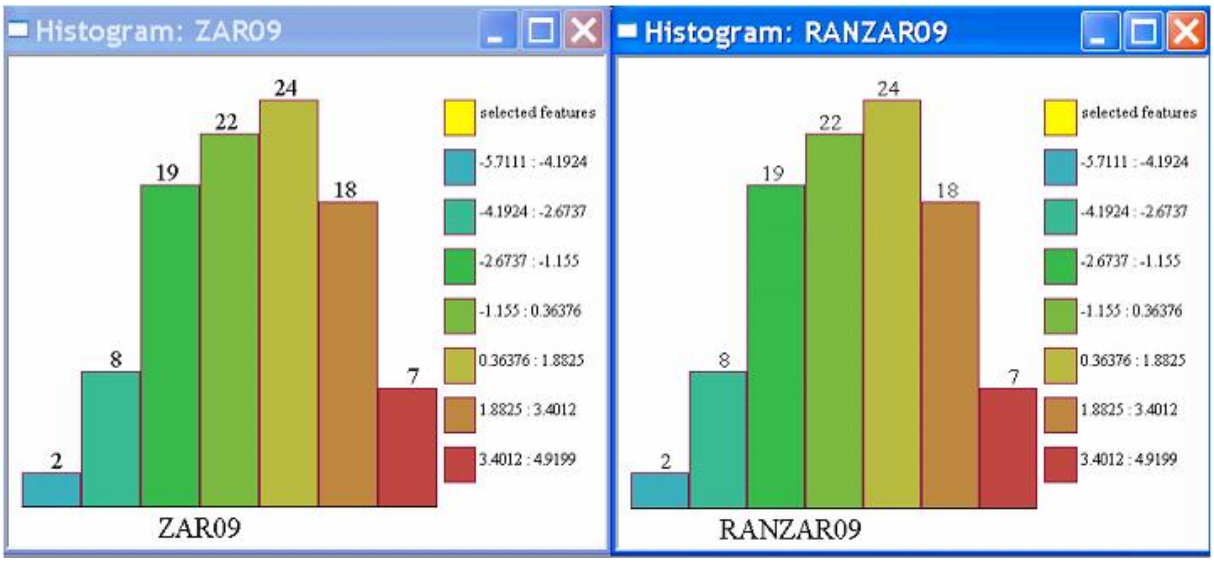


图 7.5　SAR 变量与其排序形式的柱状图

这一点可以通过链接柱状图和地图进行进一步的显示。在 zar09 的柱状图中选中(单击)最高的条，注意在另一个柱状图中的分布如何变化，如图 7.6 所示。图中相应的观察点也被高亮显示。这显示出 zar09 中最高值的位置与 ranzar09 中是不同的。

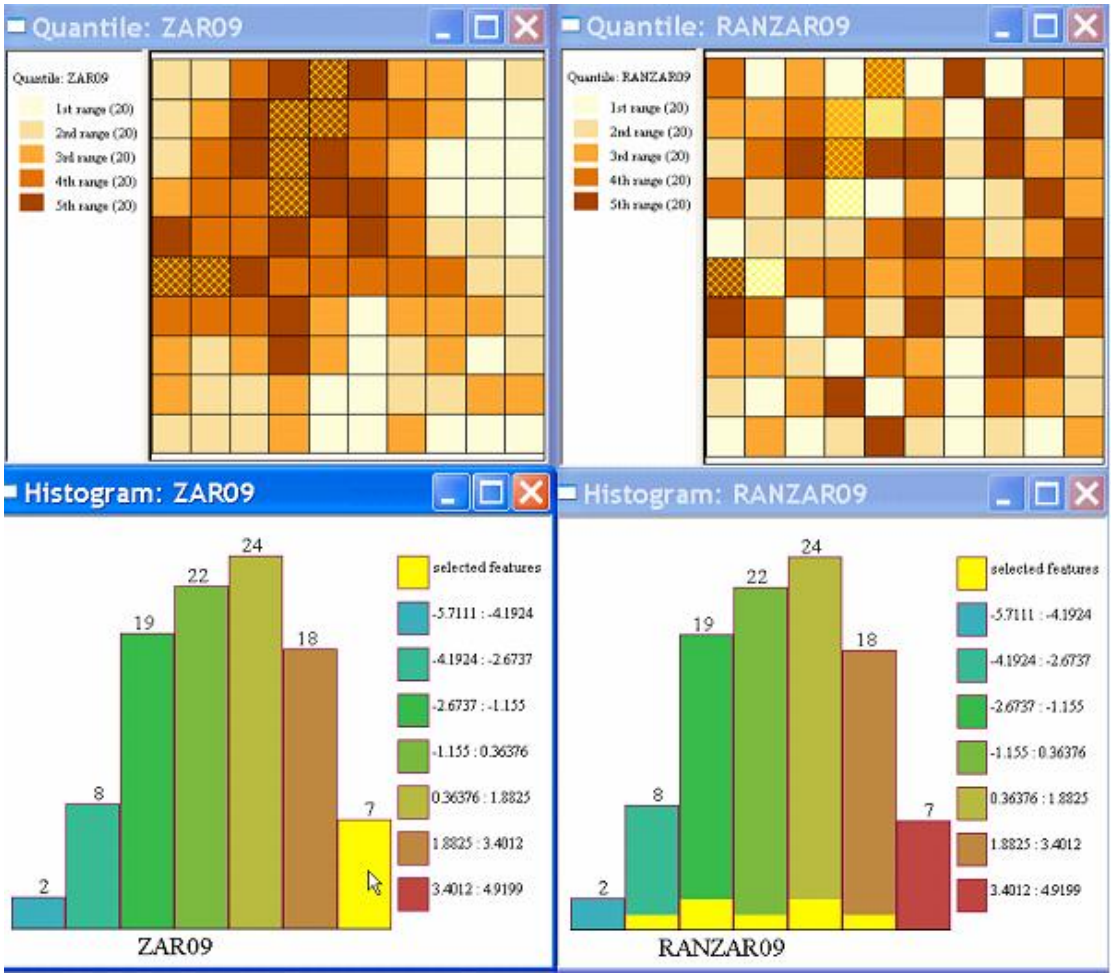


图 7.6 链接的柱状图与地图(从柱状图到地图)

可以在任何窗口进行链接。例如，在左上角地图中选择 5×5 正方形格网，如图 7.7 所示。两个

柱状图中的匹配的分布被高亮显示为黄色，显示出区域柱状图。这显示出地图中一个变量被选中的数据的位置分布。这种区域分布区别于全局分布的程度，可能暗示着存在空间性。一种被称为 spatial regimes 的特定格式，是亚区(regimes)显示明显给定变量分布的情况。例如，在图 7.7 左边的面板中，被选中的区域产生柱状图中的值(黄色高亮显示)，集中在上半部分分布。相反，在右侧面板中相同被选择的区域产生产生大致呈全局分布的值(黄色子柱状图)。这意味着 zar09 可能存在空间 regime，而 ranzar09 不存在。

可以从菜单中选择 Option>Intervals 来改变默认的分类数目 7，或在柱状图中右击，如图 7.8。选择这一选项，改变间隔数目为 12，如图 7.9。分布的黄色区域仍然与地图中被选中的数据相匹配，当它扩展到更多的类，它仍然集中分布在上半部分。

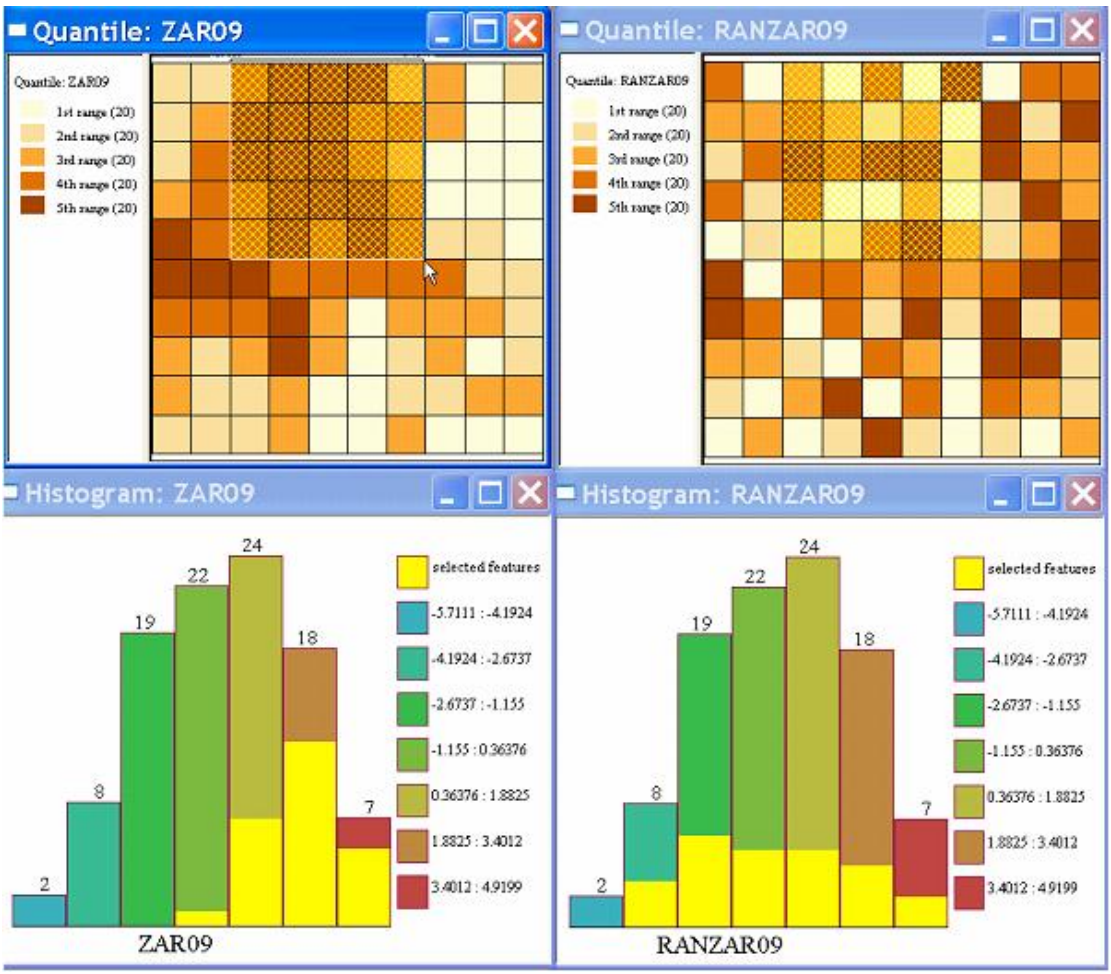


图 7.7 链接的柱状图和地图(从地图到柱状图)

图 7.8 改变柱状图类别数目

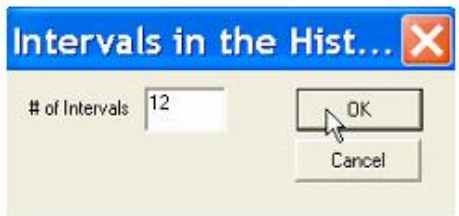


图 7.9 设置间隔为 12

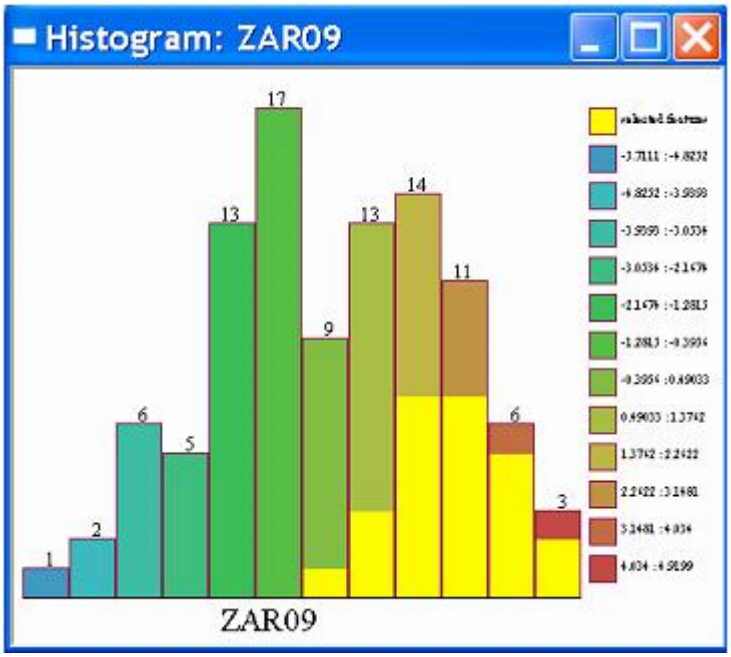


图 7.10 间隔为 12 的柱状图

7.3 链接箱图

第二种描述变量非空间分析的基本 EDA 方法是箱图(box plot)(有时称为箱图和 whisker 图)。它显示分布的中值、第一和第三分位数(在累积分布中的第 50%，25%和 70%分位点)，也显示离群值。当观测值位于给定的分位距 (75%和 25%分位值之差) 乘数，各自高于 75%或低于 25%，这时称为离群值。标准乘数为 1.5 和 3 倍分位距。GeoDa 支持这两个值。

关闭所有窗口，打开一个新的项目，stl_hom.shp 杀人案样本数据(以 FIPSNO 为关键字)。打开的屏幕会显示 78 个州的底图，如图 7.11 所示。

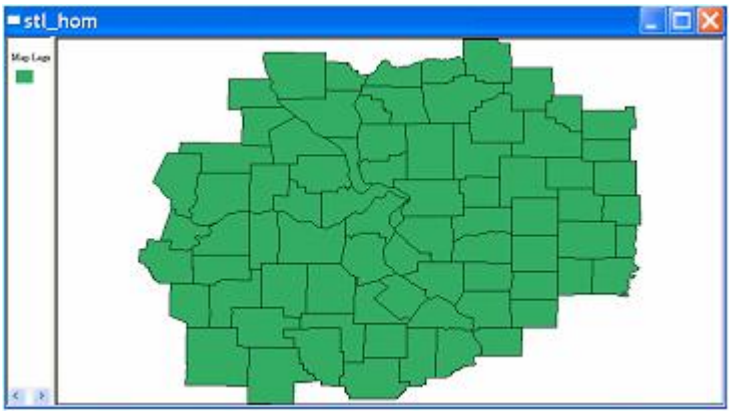


图 7.11 St.Louis 杀人案数据底图

从菜单中选择 Explore>Box Plot 启动箱图(如图 7.12)，或点击箱图工具按钮。然后在对话框中选择变量 HR8893(1988 到 1993 年杀人案发生率)，如图 7.13。点击 OK 创建箱图，图 7.14 左侧面板所示。矩形代表变量的累积分布，按值排序。右上角的值即观测点的数目。

中间的红条与中值相对应，黑色部分表明分位距(从 25%到 75%)。第一和第四分位的观测值显示为蓝色点。细线是 hinge，对应于默认的标准 1.5。这表明对于这一变量，有 6 个州被认为是离异点。

Hinge 标准决定在被分为离群值前，什么样的需要被认定为极端值。它可以从菜单选项通过选择 Option>Hinge 来改变，或在箱图中右击，如图 7.16。选择 3.0 为新的标准，观察离群值的数目

减少为 2，如图 7.15 右侧所示。

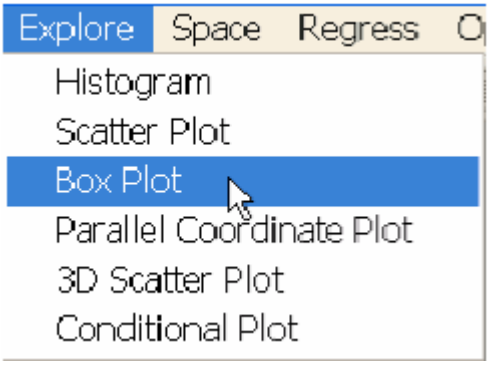


图 7.12 箱图功能

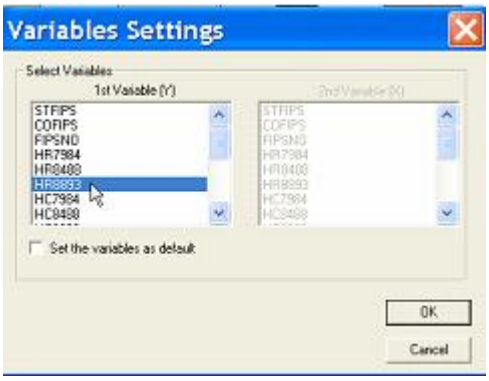


图 7.13 在箱图中选择变量

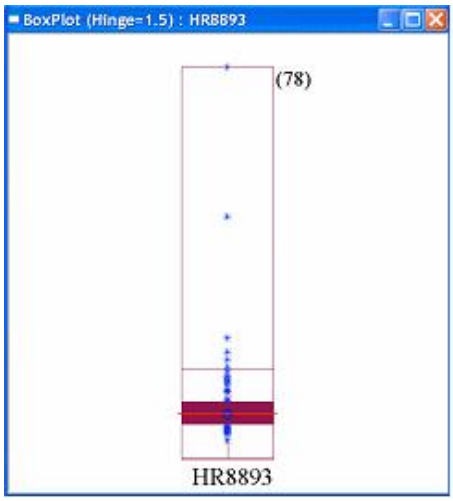


图 7.14 1.5hinge 的箱图

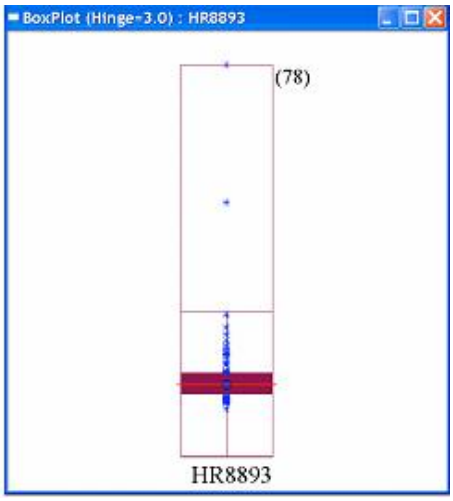


图 7.15 3.0hinge 的箱图

图 7.16 为箱图改变 hinge 标准

可以在箱图中通过单击或左拖出的选择矩形来选择某些观测点。这些选择会通过链接机制立即在所有打开的窗口中。例如，确认你有打开的 St.Louis 数据表和底图。在箱图中拖拉选择一个矩形，

选择离群值，如图 7.17 左上部分所示。注意被选中的州如何在地图及表格中被高亮显示(你可以使用 Promotion 来使表格中选中的州出现在表格顶部)。相似的，你可以在表格中选择行，看它们在箱图(或其它图)中位于哪里。

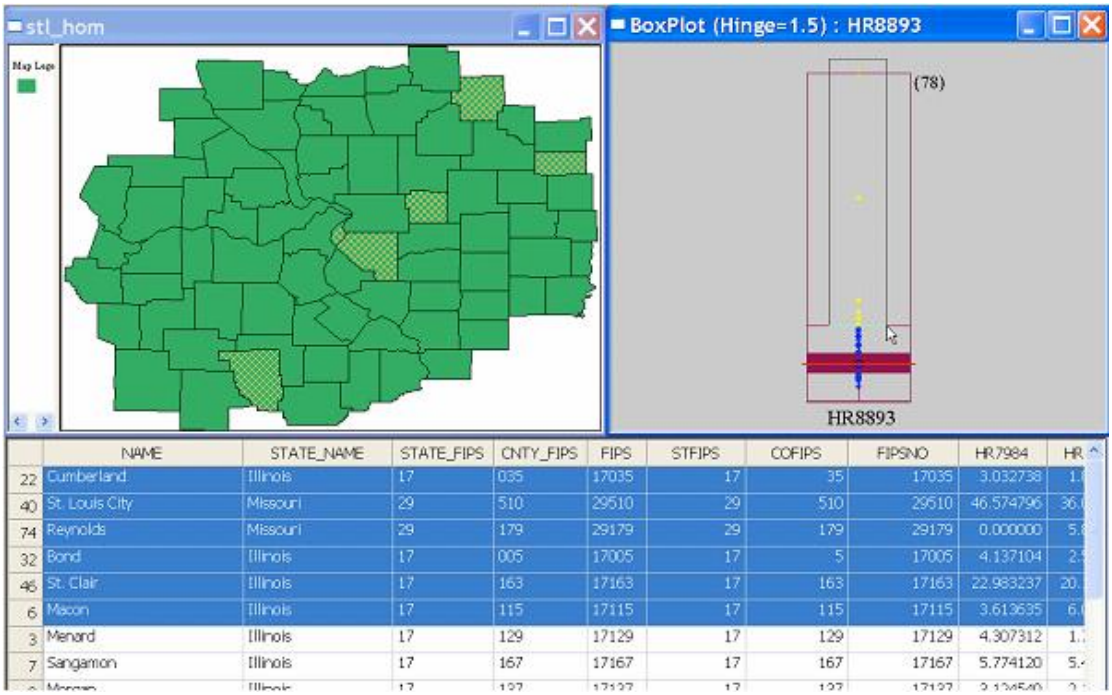


图 7.17 链接的箱图，表格和地图

在箱图中画一个小的选择矩形，按下 Control 键，然后放开。选择矩形会闪一下。这表明你开始刷光(brushing)，这是动态改变选择的方法。在箱图中缓慢上下移动刷子，注意链接地图及箱图中被选中的观测对象如何发生变化。在练习 8 中还会详细介绍刷光(brushing)。

7.4 练习

利用链接地图的 St.Louis 数据及柱状图调查杀人案发生率(HR****)三个时期的区域分布(如东部和西部，中心和边缘)。利用箱图评估始终是离群值的州。利用表格识别这些州的名字及实际的发生率(HC****)。

或者，试验其它的多边形数据来执行相同的分析，如调查 North Carolina 的 SIDS 的离群值，或 NDVI 规则格网数据的绿色(greenness)指数。

练习 8　刷光散点和地图

8.1 目标

该练习通过散点图解决两个变量相关性的可视化。GeoDa 的主要特征是刷光散点图及地图。在练习末，你将知道：

- 为两个变量创建散点图
- 将散点图转变为相关点图
- 重新计算选中观测值的散点图的斜率
- 刷光散点图
- 刷光地图

有关本操作更详细的信息见 Users Guide。

8.2 散点图

我们继续使用练习 7 中用过的数据集 St.Louis 杀人案样品数据。如果你正在对一个其它的数据进行操作，关闭所有窗口，然后打开 stl_hom，Fipsno 为关键字。从菜单中调用散点图功能，Explore>Scatter Plot（图 8.1），或点击 Scatter Plot 工具条按钮。这将启动一个变量选择对话框，如图 8.2。

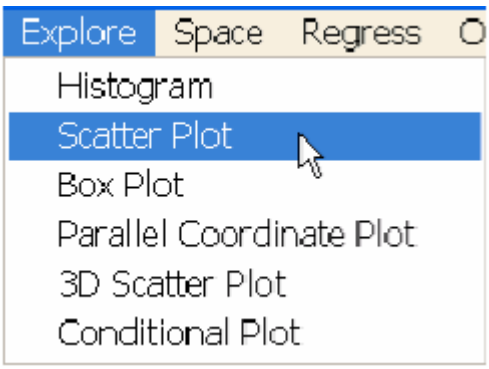


图 8.1　散点图功能

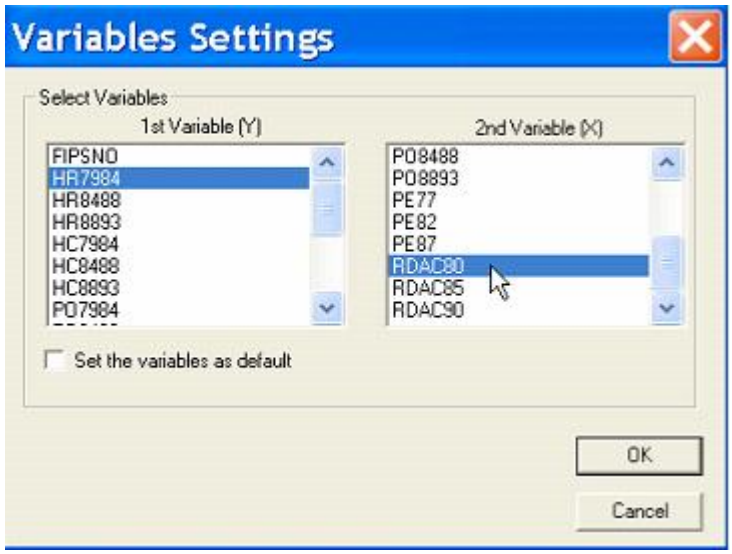


图 8.2　为散点图选择变量

在左栏选择 HR7984（1979—1984 年杀人案发生率）作为 Y 变量，右栏中选择 RDAC（由人口统计变量构建的资源缺乏指数）作为 X 变量。点击 OK 产生一个基本散点图，如图 8.3。通过散点的蓝线为最小二乘回归曲线，图表顶部为估计斜率（4.7957）。如人所料，其为正相关，高的资源缺乏程度与高的杀人案发生率间具有相关性。因为 RDAC80 有正值和负值，0 处画了一条垂线（如果所有变量为正值，就没有这条线）。

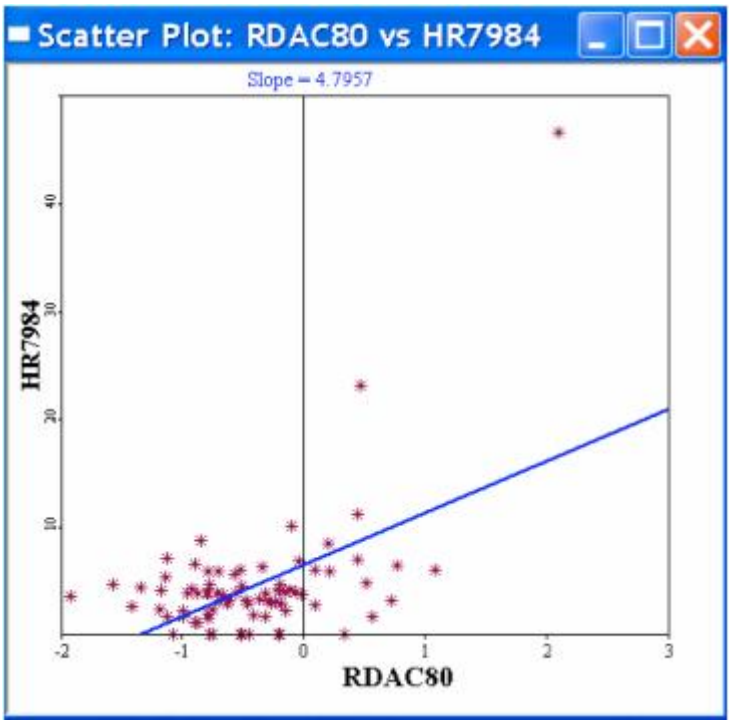


图 8.3　杀人案发生率与资源缺乏的散点图

GeoDa 中散点图有两个重要选项。选择 Options 菜单或在图中右击可以打开。如图 8.4 打开下拉菜单，选择 Scatter plot>Standardized data。这将散点图转变为相互关系图，其中的回归斜率相当于两个变量的相关系数（与默认情况下的双变量回归斜率正好相反）。

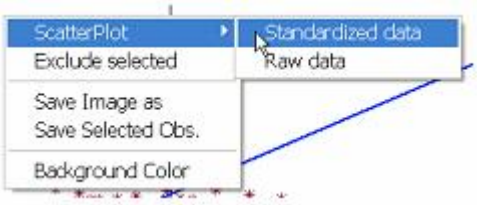


图 8.4　使用标准值选项

两个轴的变量单位调整为标准差，所以超过 2 的观测值被认为是异常值。还有，如图 8.5 所示，图形分为四个象限，可以通过类型进行其关系的定性评估：高高、低低（相对于平均值）为正相关，高低和低高为负相关。单击或拉出一个选择框选择任何一点，注意相匹配的观测点位于 St.Louis 底图上的什么地方（我们在 8.2.2 部分再谈刷光（brushing）和链接(linking)）。

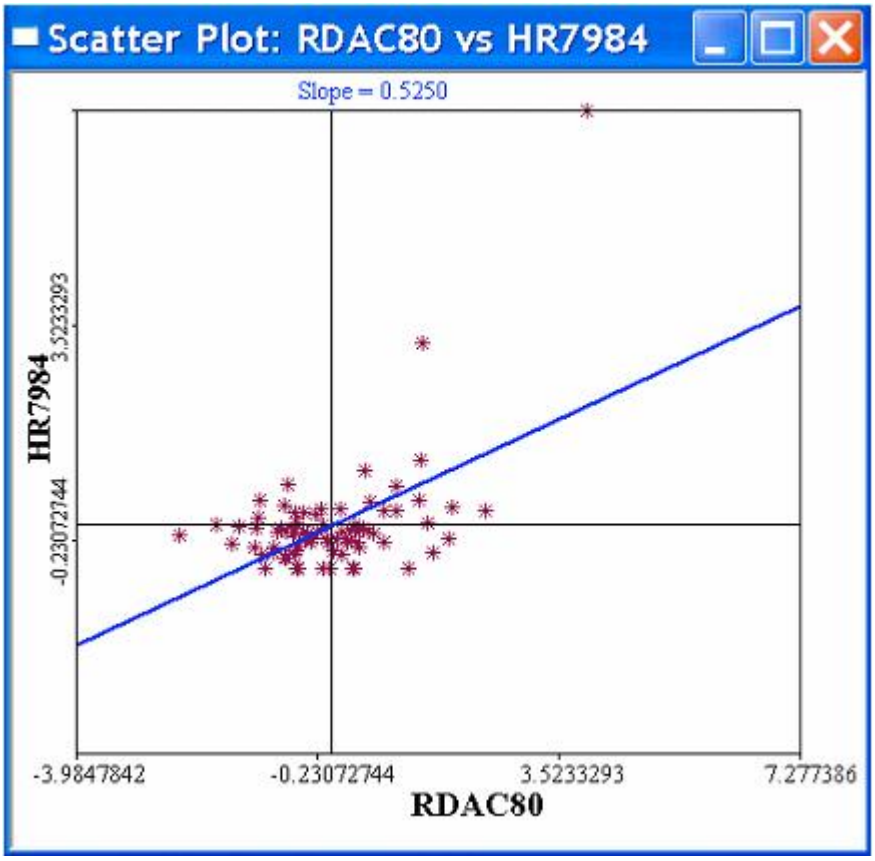


图 8.5　杀人案发生率与资源缺乏相互关系图

图顶部显示的两个变量的相关为 0.5250。在继续下一部分之前，在图中右击 Scatter plot>Raw data 返回到默认的散点图窗口。

8.2.1 剔除选择值

散点图中另一个重要选项是在剔除被选值后回归斜率动态调整。在刷光散点图时这一点非常有用。注意这一选项在相关散点图中不能正常使用，可以会产生不正确的结果，如比 1 大的相关系数。剔除选择值后，散点图的相关斜率就不再相关[1]。

通常情况下，你可以从 Options 菜单或在图中右击来选择这一选项。将会出现一具选项对话框，如图 8.6。

[1] 这一问题的原因是 GeoDa 不会重新计算标准化变量的中心，所以没有了被选数值的数据集不再是被标准化的。

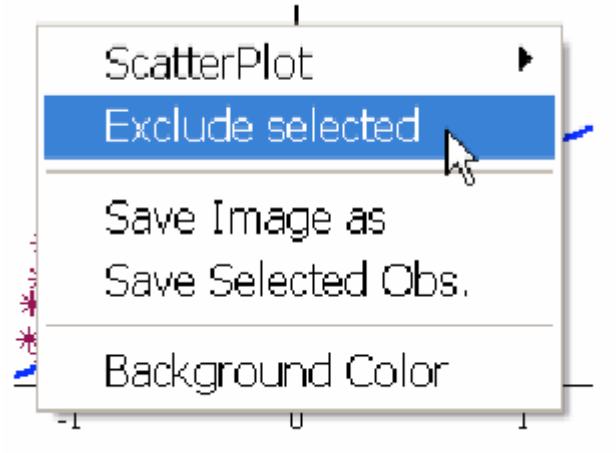


图 8.6　使用剔除被选值选项

在散点图中，选择右上角的两个点（这两个点有最高的杀人案发生率），如图 8.7 所示。注意出现了一条新的回归直线（棕色），反映了不考虑被选值的情况下这两变量的关系。新的斜率出现在图的顶部，原斜率的右边。在图 8.7 中，结果太有戏剧性（dramatic），斜率从 4.7957 直降到 0.9568。显示出这两个值（St.Louis，MO，和 St.Clair，IL（不知道 MO 和 IL 是什么意思））对斜率的强大影响[1]。

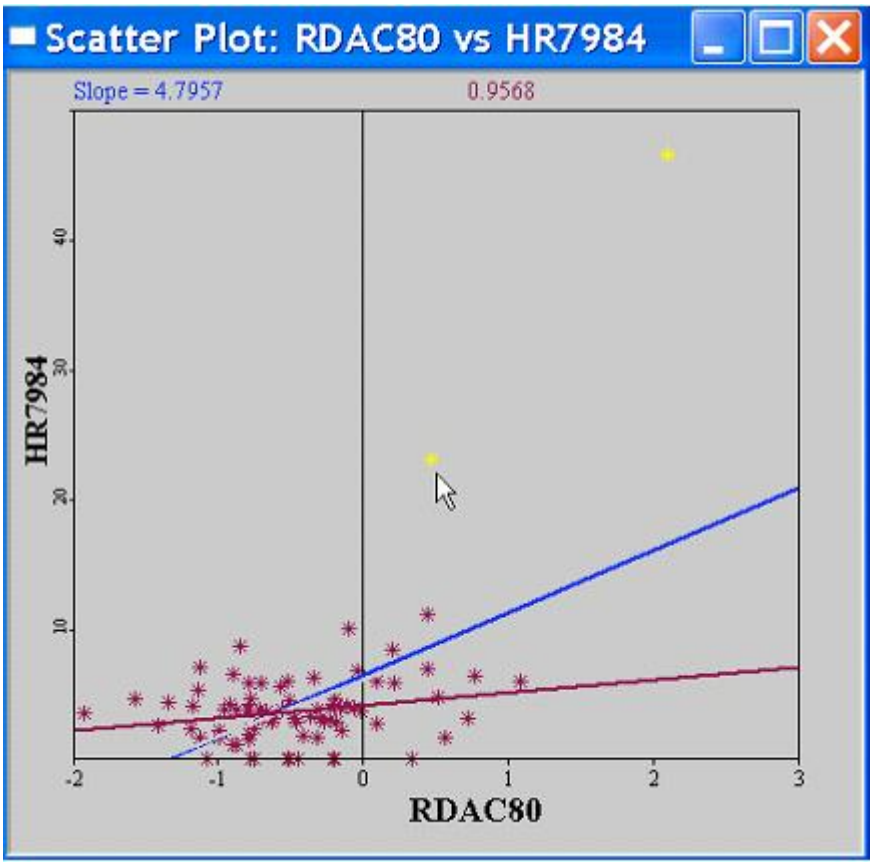


图 8.7　剔除两个值的散点图

[1] 注意有影响的观测值的分析并不具有任何参考意义，但在这一点上也是探索性的。更详细的解释请见 Messner and Anselin(2004)。

8.2.2 刷光散点图

当与被选值动态变化相联系时，剔除被选值选项成为一个非常有用的探索工具。这被称为刷光（brushing），可以在 GeoDa 的任何窗口执行。

要在散点图中开始刷光，首先确认 Exclude selected 选项是打开的。然后，创建一个小的选择矩形，按下 Ctl 键。一旦矩形开始闪烁，刷子就开始启动了，如图 8.8 所示。从现在开始，当你移动刷子（矩形），选择会发生变化：一些点变回原来的颜色，一些点变为黄色。当继续这一过程，回归直线被迅速重新计算，反映不包含当前被选值的数据集的斜率。

当刷光功能与链接（Linking）功能相联系时，就可明显看出其全部的能量。例如，在图 8.9 中，散点图与由下一时期（HR8488）的杀人案发生率的五分位图放在一起。点击“绿色”底图，利用 Map>Quantile 功能（或右击地图有相同的效果）创建这一 choropleth 图。只要地图出现，散点图中的选择会以十字阴影的形式反映出来。

刷光和动态链接的意思是当你在散点图中移动刷子，被选中的县就会变化，散点图中的斜率也会变化。这很容易扩展到其它图中，如加载一个直方图或箱图。以这种方式，在多变量意义下，开始探索变量间的联系成为可能。例如，直方图可以和不同时期杀人案发生率与资源缺乏程度相结合，调查值区域及其位置在一年中保持一致（不懂啥意思）。

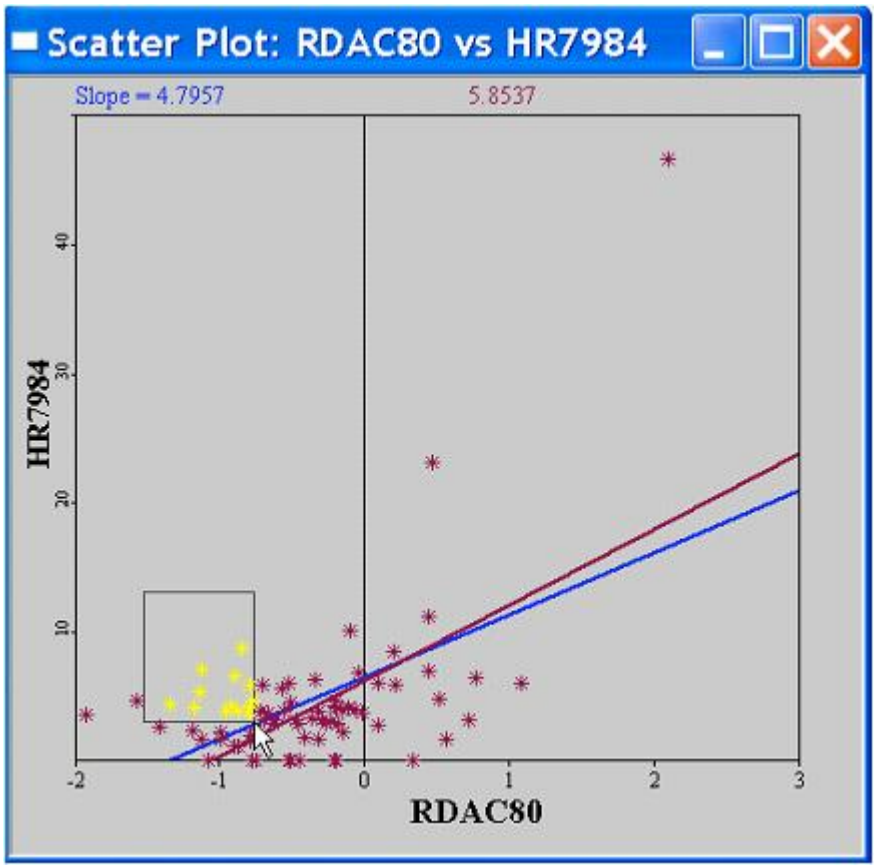


图 8.8 刷光散点图

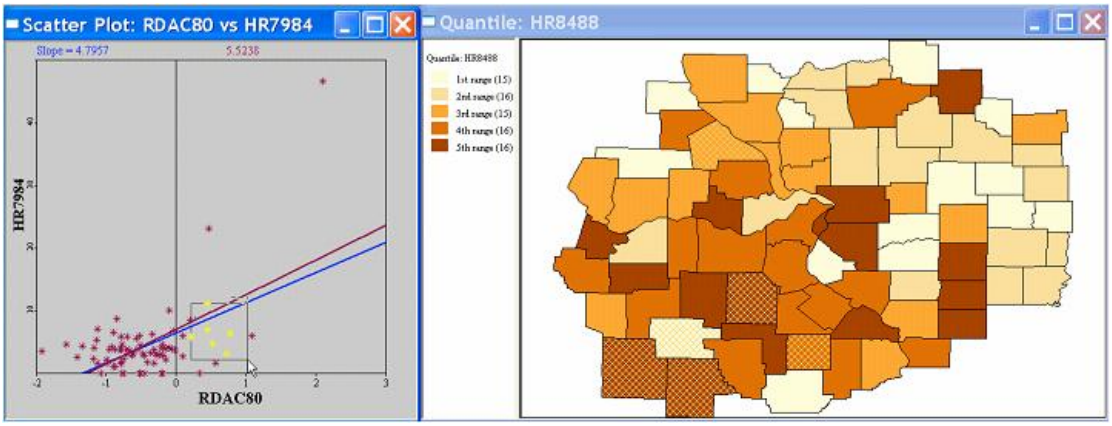


图 8.9　刷光和动态链接散点图和地图

8.3 刷光地图

在图 8.9 中，刷光功能在散点图中启动，地图中的被选择的值的变化是散点图链接的结果。相反，逻辑可以反过来。在散点图中任意地方点击关闭刷子。现在使地图成为活动窗口，以上面提到的方式建立一个刷子（按住 Ctl 键，创建一个矩形）。结果如图 8.10 所示。很快，矩形就开始闪烁，显示刷子已经准备就绪。当你在地图上移动刷子，被选中的值会变化。变化不仅发生在地图中，也会出现在散点图中。另外，当刷子在地图上移动时，散点图会迅速重新计算。

同样的，刷光也可以在任何统计图中启动，并将其选择反映在所有图表中，项目中的地图和表格。

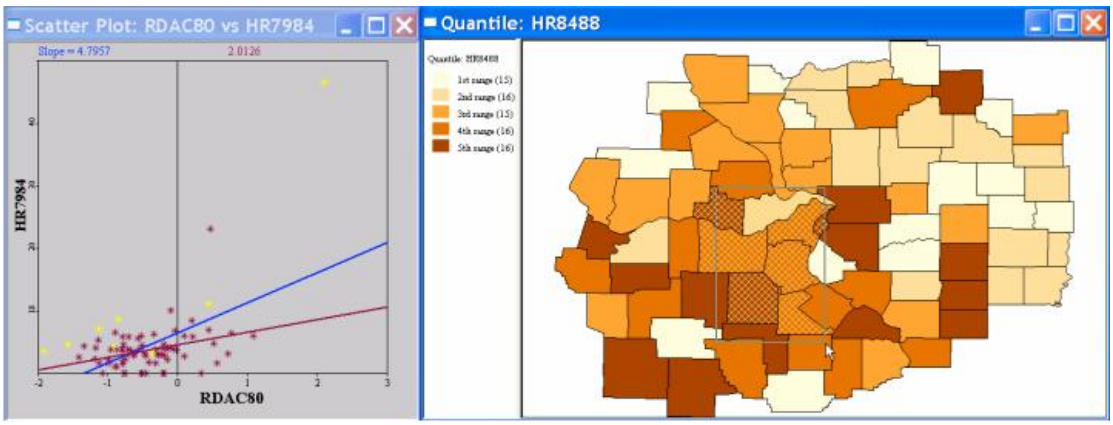


图 8.10　刷光地图

8.4 练习

在地图、散点图、直方图和箱图之间，利用刷光和动态链接继续探索杀人案发生率与资源缺乏变量之间的联系。比较其与杀人案与警用支出的联系。或者，考虑包含相同变量的 Atlanta（atl_hom.shp，关键字是 FIPSNO）和 Houston（hou_hom.shp，关键字是 FIPSNO）的样本数据。

练习 9　多变量 EDA 基础

9.1 目标

本练习通过散点图矩阵和平行坐标图来解决多变量之间联系的可视化问题。

在本练习末，你将会知道：

- 将散点图和其它图表布置成为一个散点图矩阵
- 刷光散点图矩阵
- 创建一先进平行坐标图
- 刷光平行坐标图

有关该操作更详细的解释请见 Release Notes。

9.2 散点图矩阵

我们将会探索工 POLICE 样本数据集(输入 police.shp 文件名，关键字是 FIPSNO)82 个 Mississippi 县的警用支出、犯罪和失业之间的联系。

开始的底图应该如图 9.1。

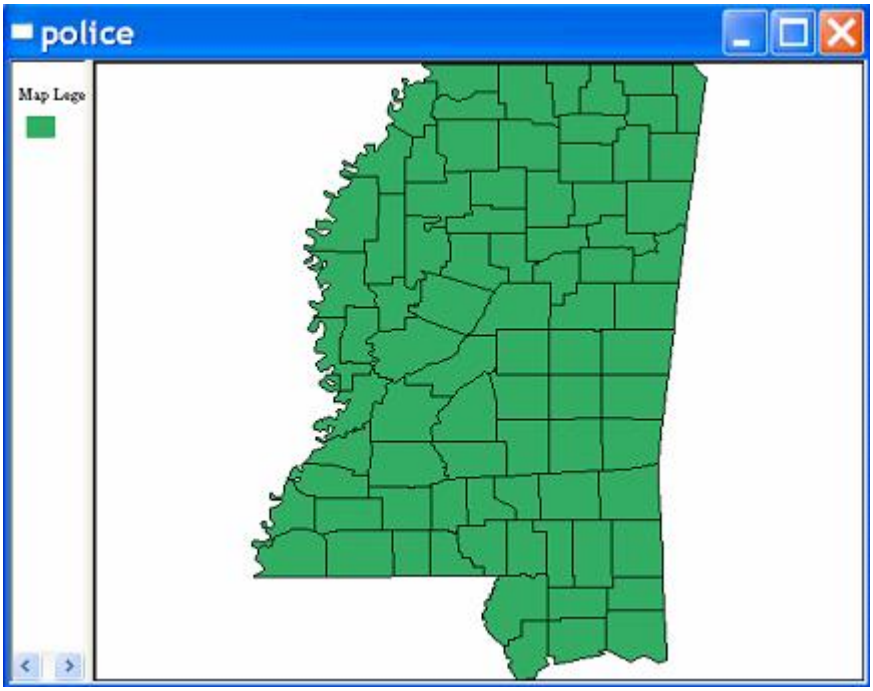


图 9.1 Mississippi 县警用支出数据底图

需要考虑的第一种方法是散点图矩阵。即使本身(per se)GeoDa 不包括这个功能，为刷光和链接建立这一方法也是可以的(尽管有些麻烦(tedious))。矩阵由一系列成对的散点图构成，布置成为所有一行的散点图的 Y 轴都是相同的。诊断块(blocks)是空的，或用于单变量统计图表。

例如，为 police 变量(Map>Quantile，分为 5 类，见练习 2)创建一幅五分位图。将图例和地图

之间的垂直分界线移到左边，图例就会消失，如图 9.2。变量 crime 和 unemp 也重复这一过程。然后，为变量 police 和 crime 创建两个双变量散点图(见 8.2 部分的详细指导)，每个变量轮流位于 Y 轴。布置散点图和两个相匹配的五分位图到 2×2 的矩阵中，如图 9.3。

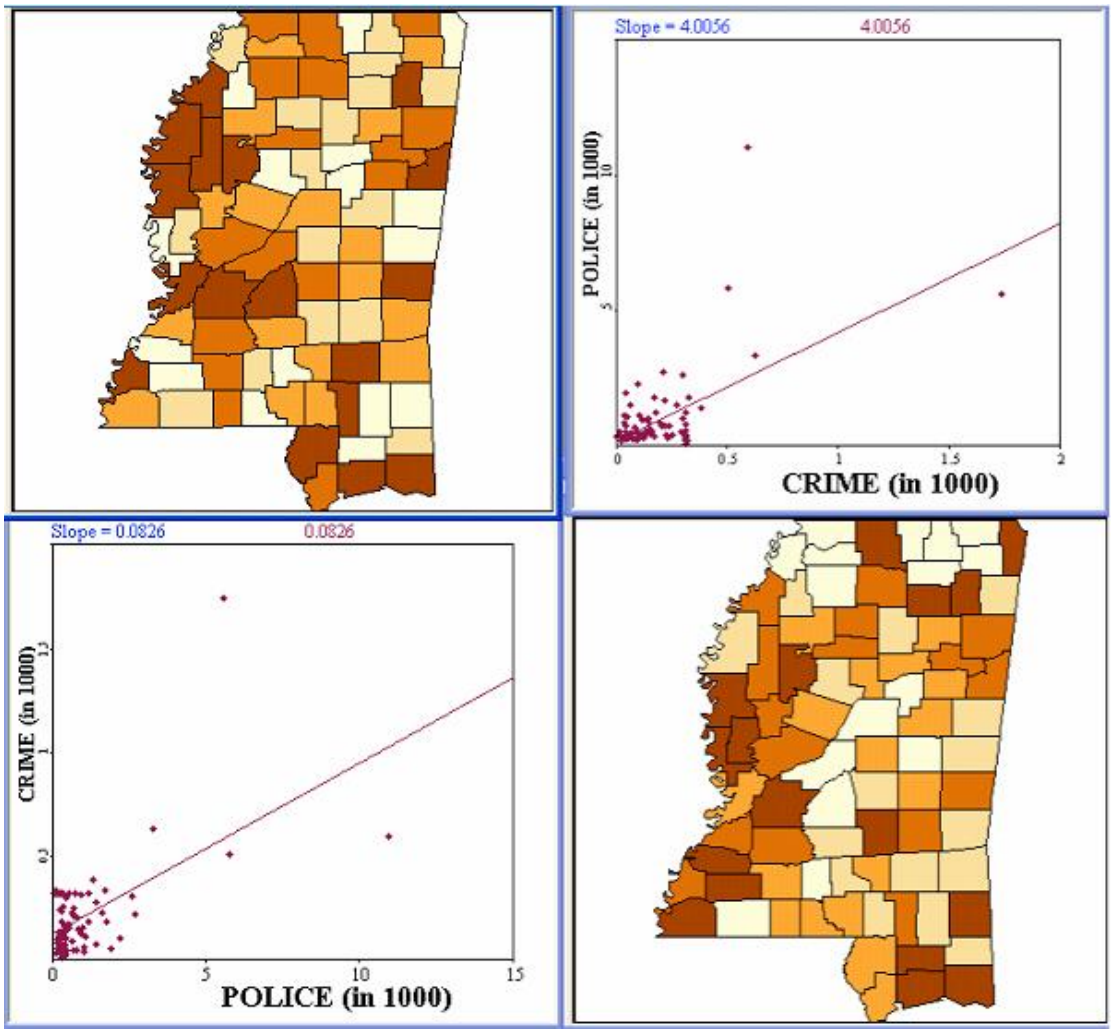


图 9.3 2×2 散点图矩阵(police，crime)

为剩下的 police 和 unemp，unemp 和 crime 之间的散点图，继续这一操作(这是一个乏味的过程)。确信在每一散点图中 Exclude selected 选项是打开的(因为这一选择是对特定图表的，必须单独为第一个单独设定)。为便于浏览被选点，你也许希望背景颜色为灰色或改变选中状态的颜色。重新布置各种图表之后，散点图矩阵将会看起来如图 9.4。

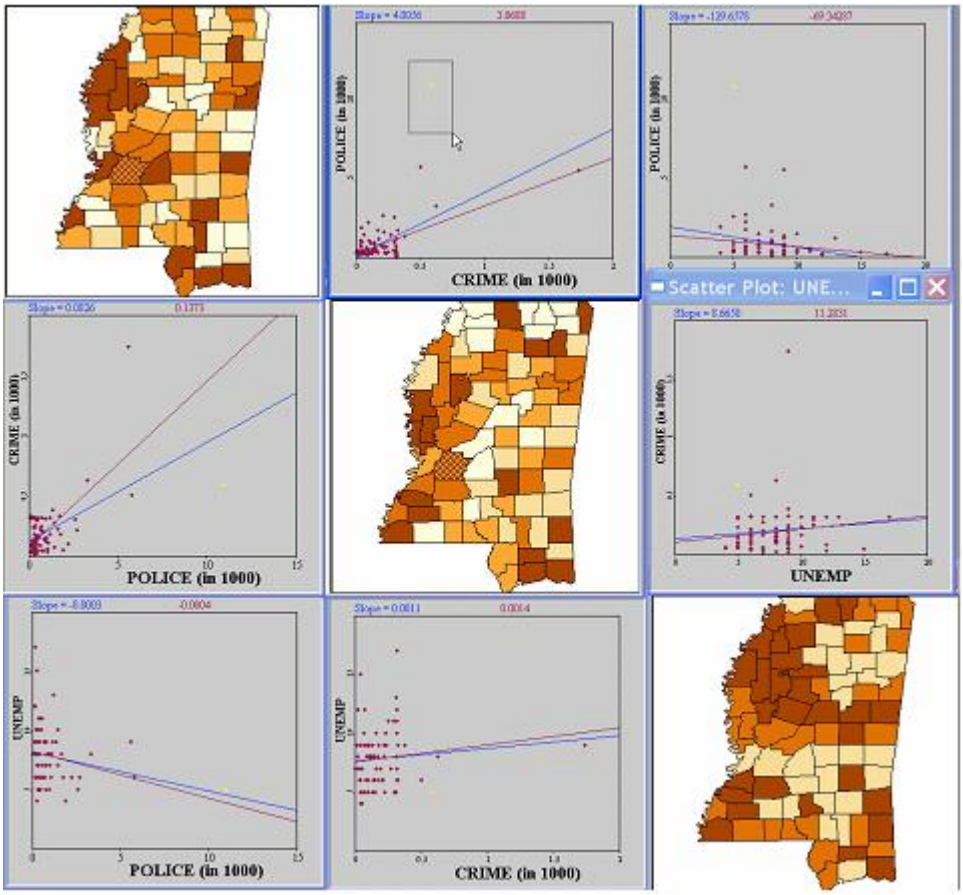


图 9.4 刷光散点图矩阵

你现在可以通过仔细检查对(pairwise)斜率、选择兴趣点、刷光散点图来探索变量间的多变量联系。如在图 9.4 中，警用支出与犯罪是正相关的(如人所料)，但与失业与负相关。另一方面，犯罪与失业成正相关，暗示着在这些变量之间存在着更复杂的相互作用，是不能用简单的两两联系来显示的。通过刷光(一次只能有一个刷子可用)可以评估斜率对某一观测值的敏感性。例如，在图 9.4 中，在 police-crime 散点图中刷光最高警用支出的点，将出现低的多的斜率。因为所有的图表是链接的，被选中的观测点也在诊断矩阵的其它三个图中被高亮显示。

9.3 平行坐标图

散点图矩阵的一个替代者平行坐标图(PCP)。多变量比较中要考虑的每一个变量都成为图中一个平行坐标轴(相对于散点图的轴是正交的)。在每一个轴上，变量的观测值是从最低(左)到最高(右)。结果，多变量被表示成一系列的线段，连接在每一个轴上相应的位置。这些线段是与多变量(多维)散点图中的点相对应的。

从菜单(图 9.5)中选择 Explore>Parallel Coordinate Plot 或单击工具按钮，打开 PCP。出现一个变量选项对话框，如图 9.6。

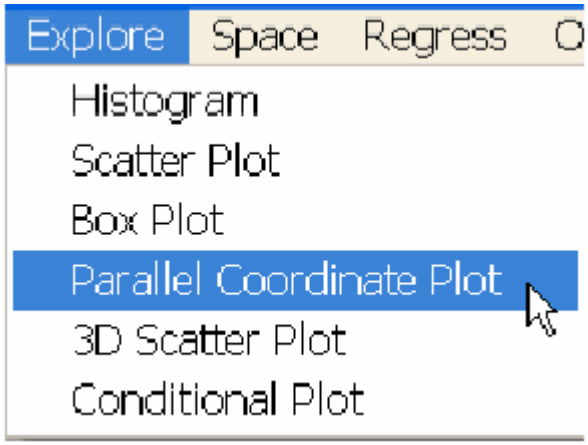


图 9.5 平行坐标功能

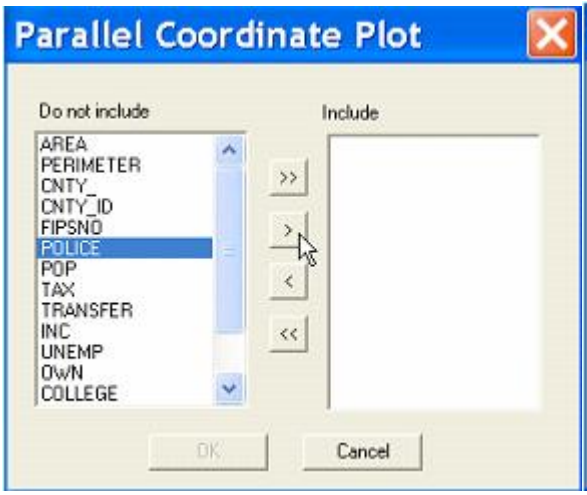


图 9.6 PCP 变量选择

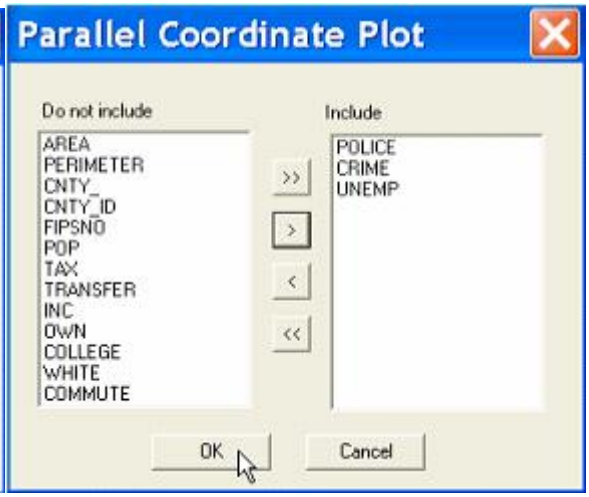


图 9.7 PCP 中选中的变量

例如，选择散点图中所用的三个相同的变量(police，crime，unemp)，单击>按钮，将它们从左边移动到右边(如图 9.6)。当这些相关变量都在右栏中，点击 OK(图 9.7)创建图。结果如图 9.8。

你可以在相应直线上点击或使用选择多边形(任何横切线将会被选中)，来选择单独的观测值。例如，点击 crime 的直线的最高值部分。它会变成为黄色(如图 9.8)，显示被选中。为更便于识别被选中的直线，图的背景颜色由默认的白色变为灰色。

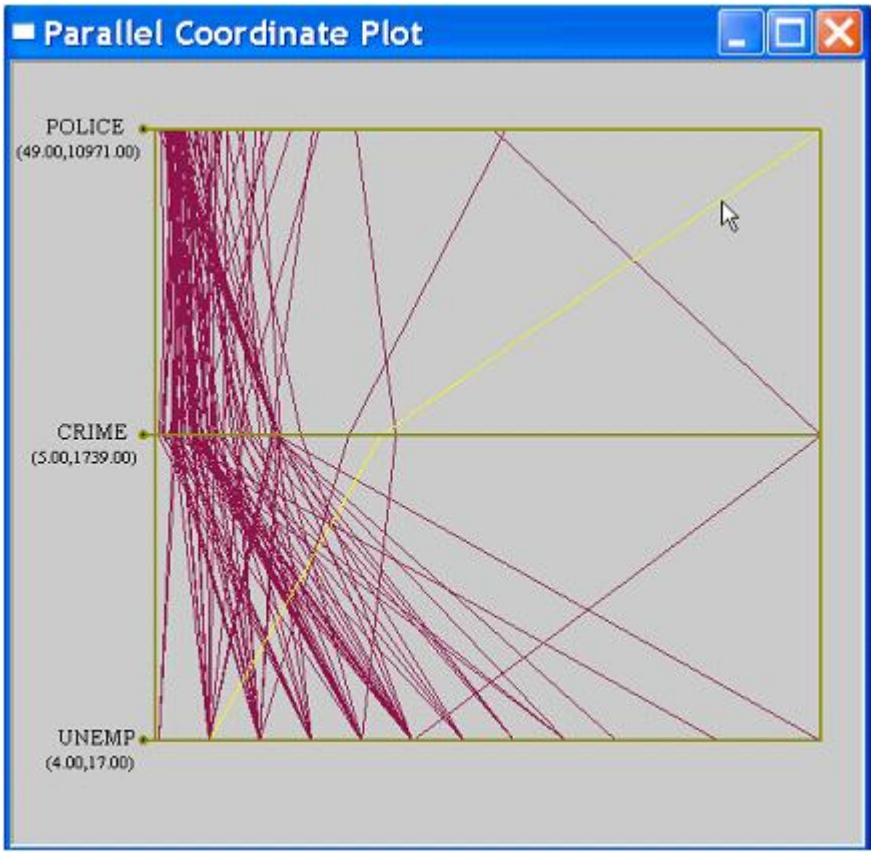


图 9.8 平行坐标图(police，crime，unemp)

被选中的部分是 Hinds 县，包括州府 Jackson，MS。注意这一位置不但警用支出高(轴的右端)，犯罪高(尽管就值大小而言，分布中心的值比警用支出高，为负斜率)，但就失业而言又比较低(在轴的左边)。

重要的是你要记住，每一个变量已经重新调整过，最小值在左端，最大值在右边。随着积极性的增加，观测值从左向右排序，相对于观测值变程(最大值与最小值之差)进行定位。

你可以改变轴的顺序来更详细地关注两个变量的联系。在图中 unemp 旁边小点上单击，向上移动，如图 9.9。当到达中间轴的位置，放开：crime 和 unemp 的变量轴会交换位置，如图 9.10。

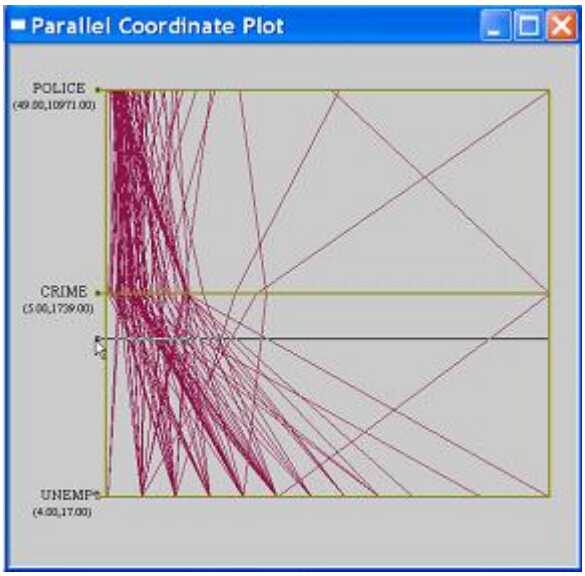


图 9.9 在 PCP 中移动轴

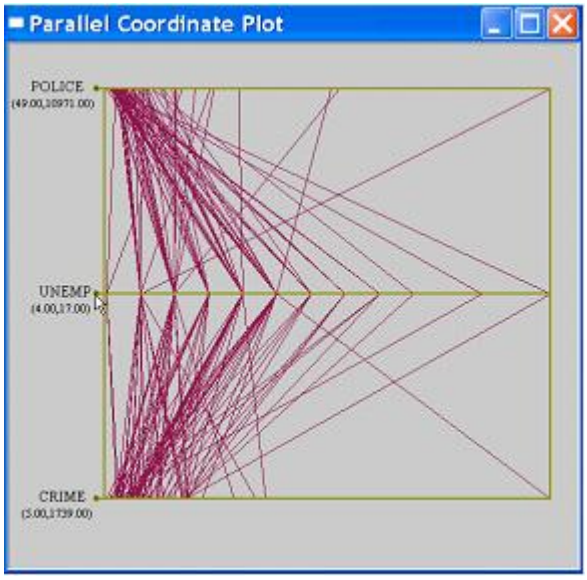


图 9.10 轴移动后的 PCP

PCP 的一个重要应用在于识别多变量空间中的观测值的丛聚。在图中以相同的直线 signature 来反映。刷光图来查找表现出相同模式的线段，查找它们在地图中的位置。你也可以得到数据表中县的名称(你需要在 Table 中使用 Promotion 选项来更容易地找到被选中的观测值)。例如，在图 9.11 中，选择刷沿 police 轴移动。当一些线段服从相同的模式，许多线段却并不相同。沿着其它的轴刷光(如果希望的话，交换轴的位置)来探索潜在的丛聚。

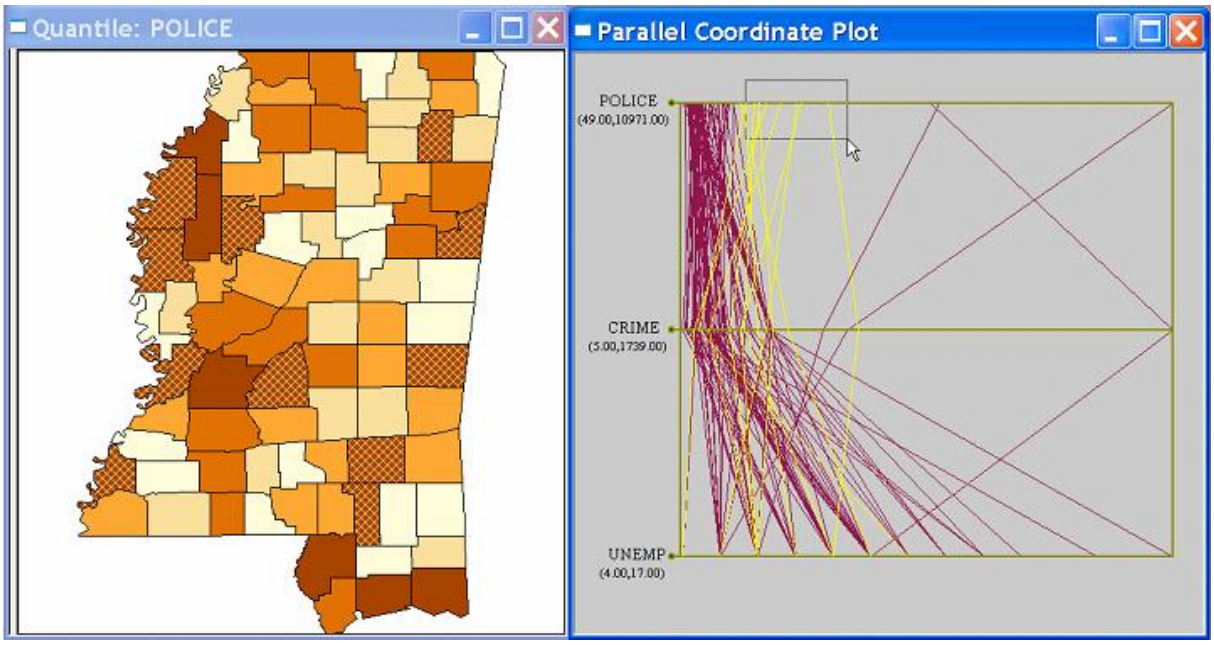


图 9.11 刷光平行坐标图

因为 PCP 与项目中其它图表都是链接的，可以评估多变量丛聚相对应的空间丛聚的程度。在图 9.11 中，PCP 与警用支出五分位图放在一起，PCP 中被选中(刷光)的观测值也在地图中高亮显示。相反，试着通过在地图中创建刷子，评估地图中亚区相对于多变量丛聚的程度。

9.4 练习

你可以利用 POLICE 数据集中的警用支出、犯罪和其它变量，进一步试验这些方法。将你的想法与 Kelejian and Robinsonr 回归分析结果相比较。

或者，重新打开 St.Louis 杀人案数据，或 Atlanta and Houston 杀人案数据，探索练习 8 中详细谈到的杀人率、资源缺乏(RDAC**)和警用支出(PE**)的联系。

练习 10　高级多变量 EDA

10.1 目标

本练习通过条件图和三维散点图的更高级的方法来探索变量间的多变量关系。

在本练习末，你将知道：

- 创建条件直方图，箱图和散点图
- 在条件图中改变条件间距
- 创建三维散点图
- 缩放和旋转 3D 散点图
- 在 3D 散点图中选择观测点
- 刷光 3D 散点图

更详细的信息请参考 Release Note 第 33-43 页。

10.2 条件图

我们继续使用练习 9 中的 POLICE 数据来探索多变量模式。之前，确保县质心已加入到数据表中(按照 6.2.1 的指导)。在本例中，我们将这些质心称为 XC00 和 YC00。

一个条件图由 9 个小图组成，每一个都是观测数据的一个子集。以两个其它变量来得到这些子集。每个变量分为三个部分来定义子集(共 9 对间距)。

点击相关工具按钮或从菜单选择 Explore>Conditional Plot，开始条件图制作，如图 10.1。将启动一个对话框，进行四种不同类型图的选择，如图 10.2 所示。两个图是单变量的(直方图和箱图，练习 7 中已涉及)，一个是双变量(散点图，练习 8 中已涉及)，一个是地图[1]。选择 Scatter Plot 旁的单选按钮，然后点击 OK。

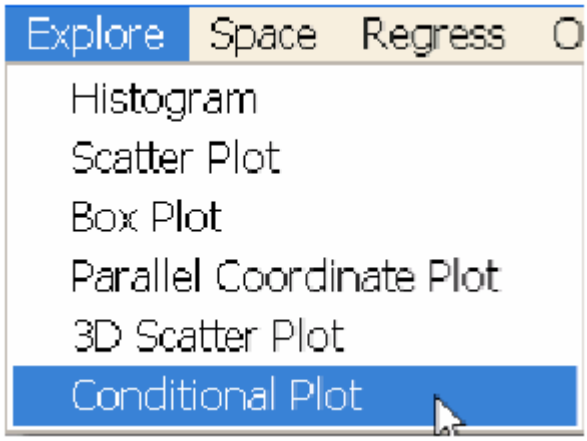
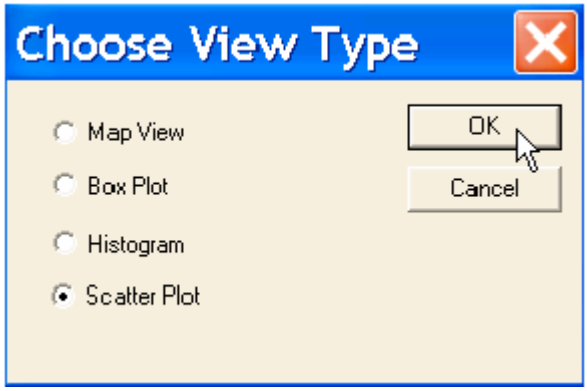


图 10.1 条件图功能

[1] 条件地图将在练习 12 中讲解。



10.2 条件散点图选项

然后，你需要指定条件变量和兴趣变量。在变量选择对话框，从下拉列表中选择一个变量，通过点击>按钮，将其移到右边的文本框，如图 10.3。特别地，为第一个条件变量(对话框中的 X 变量)选择 XC00，第二个条件变量(Y 变量)选择 YC00，POLICE 为散点图中的 Y 轴变量(变量 1)，CRIME 为 X 轴变量(变量 2)1。全面的设置如图 10.4。点击 OK 创建条件散点图，如图 10.5 所示。

更近的考虑这幅图。第一个条件变量位于水平轴上，它东西方向县的位置。根据县的质心的 X 坐标(XC00)是否落在-91.44603 到-90.37621 之间，-90.37621 到-89.30639 之间，-89.30639 到-88.2367 之间。县被分为三个“大箱子”。第二个条件变量位于纵轴的变量，对应于南北方向，也分为三个间隔。结果，这九个小图对应于肥的子集，根据从西南到东北的地理位置来安排2。

散点图暗示在警用支出和犯罪的回归斜率存在很大的区域差异。注意仍然需要探索和谨慎地解释。由于每一个图中的观测点数量差异，估计斜率相关系数的精度也会不同。这将会在更严密的比较中来考虑，如差异(variance)分析。然而，图中显示了州府有关这一双变量关系的强烈效果(左栏的中间一个图)，产生最大的斜率。在两幅图中，斜率甚至是负的。

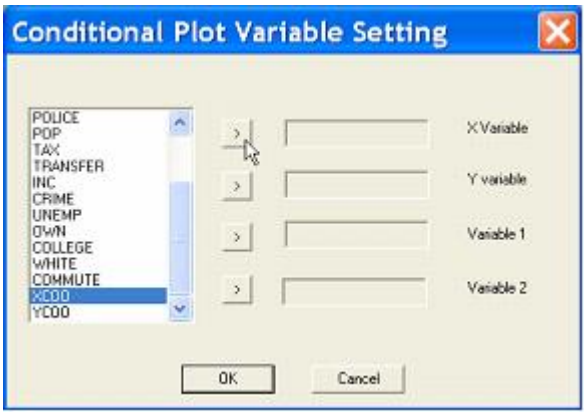


图 10.3 条件散点图变量选择

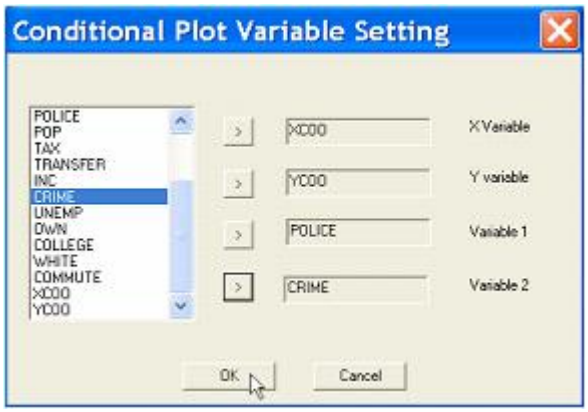


图 10.4 在条件散点图中被选择的变量

[1] 其它的条件图的对话框也是相同的，除了条件变量，只有一个变量被指定。
[2] 条件变量没有地理位置，但可以是兴趣维。

图 10.5 条件散点图

可以移动“手柄”到一边或上下移动来调整类型。手柄是现个间隔条上的两个小圆圈。通过将最东边和最北边的类型折叠在一起，可以将矩阵转为 2×2 的格式，拖动最右边的手柄到，如图 10.6。同样的，拖动最上边的手柄到顶部。这个 2×2 的的分类仍然暗示最西边的县的差异。实验移动手柄，来好好感觉一下随着陈福今定义的改变，图是如何变化的。也试着利用不同的变量(如 tax 和 white)作为条件变量。

10.3 3-D 散点图

我们探索多变量联系的最终方法是三维散点(或立方体)。关闭条件散点图，点击工具条或从菜单选择 Explore>3D Scatter Plot 打开 3D 散点图界面(如图 10.7)。

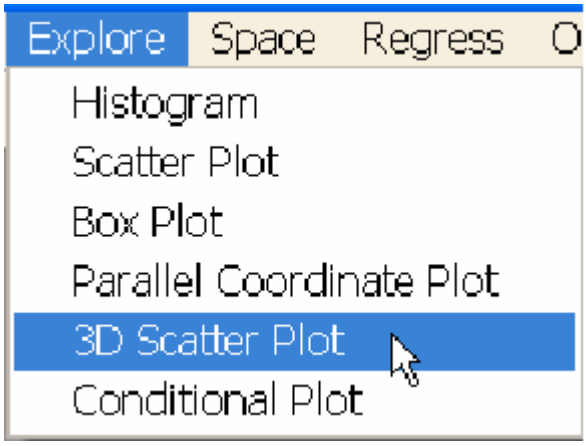


图 10.7 3 维散点图功能

进入变量选择对话框，如图 10.8。在下拉列表中为 X 变量选择 CRIME，为 Y 变量选择 UNEMP，为 Z 变量选择 POLICE，如图 10.9。注意变量的顺序并不重要，因为数据立方体可能方便的旋转(如将 X 轴由水平转为垂直)。最后点击 OK 产生一个最初的 3D 视图，如图 10.10。

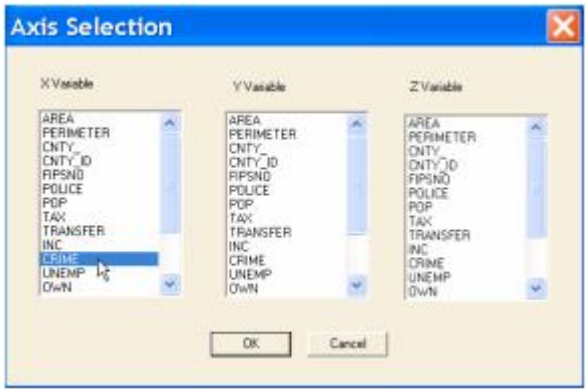


图 10.8 3D 散点图变量选择

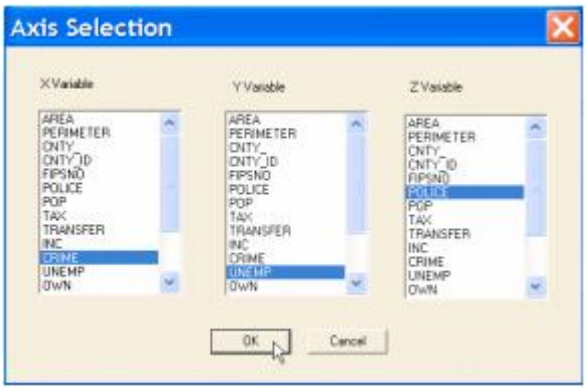


图 10.9 在 3D 散点图中被选择的变量

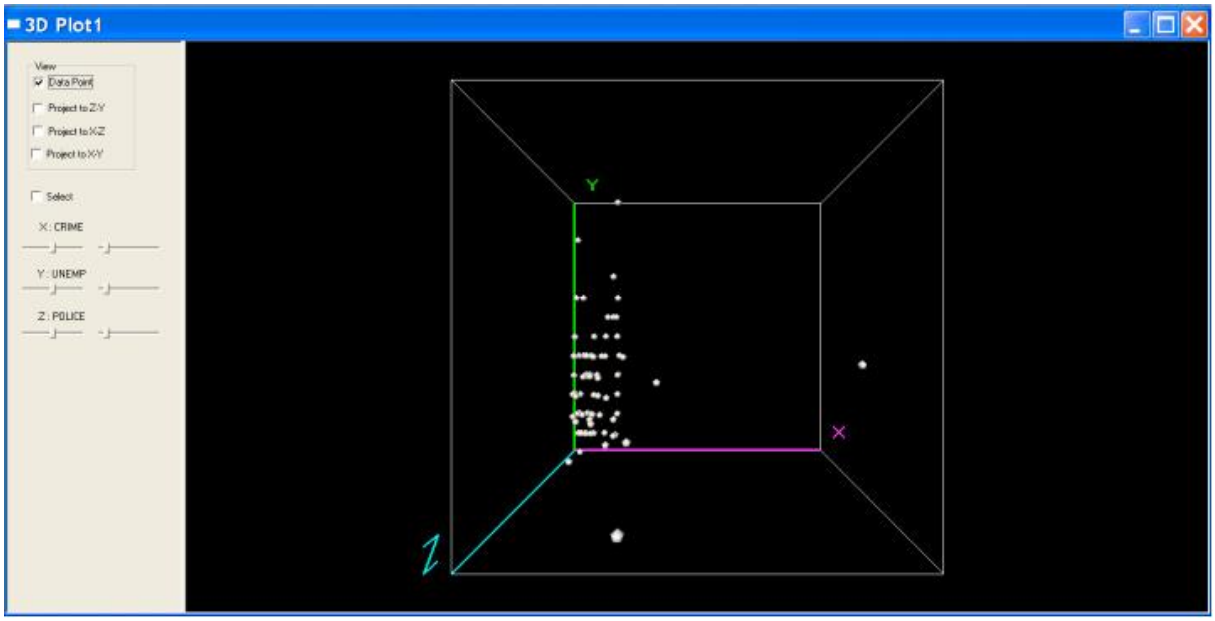


图 10.10 3D 散点图(police，crime，unemp)

还有许多方法对 3D 图进行操作。如，在图中任何部分点击，移动鼠标可以旋转图。右击可以缩放图。更重要的是，，在窗口左边面板中的一些可用的选项。在面板顶部是 3 个多选按钮，使 3 维点