文章编号：1007-4619 (2002) 02-0102-06

用 ESDA 技术从 GIS 数据库中发现知识

马荣华，黄杏元，朱传耿
(南京大学地理信息系统与遥感研究所，江苏 南京 210093)

摘 要： 从 GIS 数据库中可发现许多知识，这在 GIS 界已引起了相当的重视。ESDA 技术注重研究数据的空间依赖与空间异质性，在知识发现中用于选取感兴趣的数据子集，并可初步发现隐含在数据中的某些特征和规律。一些标准全程和局部空间统计包括 Moran I，Geary C，G 统计以及 LISA 等是 ESDA 技术的基本核心内容之一。在目前专家系统尚不成熟的条件下，充分利用 GIS 的可视化和空间分析技术是实现 ESDA 技术与 GIS 紧密完全结合的关键。实例说明用 ESDA 技术结合其它相关领域的知识从 GIS 数据库中发现知识的方法是可行的。

关键词： ESDA；GIS；数据库；知识
中图分类号： P208/TP391 **文献标识码**： A

1 引 言

数据库中存储有大量数据和信息，从中导出所隐含的知识并将这些知识作用于现有的数据以得到新的知识和数据，是当前人工智能及知识和数据工程领域的研究热点[1]。从 GIS 数据库中发掘有用的隐含知识将会提高 GIS 的应用水平，并对建立智能化的 GIS 起到极大的促进作用，在 GIS 界已引起了相当的重视[1—4]。但其概念已大大超越了 GIS，在计算机科学中也被列为新的探索领域[4]。从 GIS 数据库可以发现的主要知识类型包括：普遍的几何知识、空间分布规律、空间关联规律、空间聚类规则、空间特征规则、空间区分规则、空间演变规则和面向对象的知识[2]。其中普遍的几何知识和空间分布规律可用数量地理的有关理论和技术来发掘，这一点被普遍认识，但深入探讨不够。虽然传统的数量地理理论在 GIS 革命的潮流中，研究领域迅速向 GIS 理论与应用方面拓展，突出了数量地理学家注重空间分析的地理学特色[5]，但其在数据结果的表达方面缺乏有效的可视化手段，或者说可视化手段非常单一；另一方面，数据挖掘和知识发现对数据的可视化提出了更高的要求，这给传统的数量地理带来了冲击，并推动了数量地理的发展，必将在数据挖掘和知识发现领域发挥越来越重要的作用。数据挖掘(Data Mining，简称 DM)和知识发现(Knowledge Discovery from Database，简称 KDD)都是从数据库中提出隐含的、感兴趣的、高水平的模式，其本质是一样的[2]，因此本文视这两个概念相同。ESDA (Exploratory Spatial Data Analysis，探索性的空间数据分析)技术在数量地理学的发展过程中逐渐形成了一种成熟的、应用面甚广的技术，也是空间统计的一门前沿性技术，有关的应用研究很多[6,7]，它注重研究数据的空间依赖与空间异质性，在知识发现中用于选取感兴趣的数据子集，并可初步发现隐含在数据中的某些特征和规律[8]。

2 ESDA 的有关理论

ESDA 是指利用统计学原理和图形图表相结合对空间信息的性质进行分析、鉴别，用以引导确定性模型的结构和解法，本质上是一种“数据驱动”的分析方法[8]。它注重研究数据的空间依赖与空间异质性，即描述空间分布、揭示空间联系的结构，给出空间异质的不同形式，发现奇异观测值。所使用的方法主要有：直方图、频率分布表、概括性统计量(平均

收稿日期：2000-12-25；**修订日期**：2001-02-26
基金项目：江苏省教育委员会高校科研项目(编号：99SJB790006)和高等学校博士点专项科研基金(编号：20010284011)。

作者简介：马荣华(1972—)，男，山东临沂人，南京大学城市与资源学系博士生，主要从事地理信息系统软件设计开发与数据处理研究，发表论文 10 余篇。

值、最大值、最小值、均方差、倾斜(skewness)系数等)、散点图、方差图(方差云图和异质方差图)、Moran I, G 统计量和局部空间统计量(Local Spatial Statistics, LISA)以及空间自回归模型等,表现结果的手段除传统的图形图表外,还可与 GIS 相结合,利用 GIS 的可视化技术,把相关结果表示到基础底图上,增强直观效果。ESDA 技术应用的条件是零假设(Null Hypothesis)即空间不相关假设。

2.1　空间权重矩阵

空间权重矩阵的定义是空间统计学与传统统计学的重要区别之一,是利用 ESDA 技术进行空间探索分析的前提和基础。其目的是定义空间对象的相互邻接关系,抓住 GIS 数据库中有关数据的空间联系。矩阵如(1)式。

$$\begin{pmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ W_{m1} & W_{m2} & \cdots & W_{mn} \end{pmatrix} \tag{1}$$

上述权重矩阵的相邻规则:

$$W_{ij} = \begin{cases} 1 & \text{位置 } i \text{ 与位置 } j \text{ 相邻;} \\ 0 & \text{位置 } i \text{ 与位置 } j \text{ 不相邻。} \end{cases} \tag{2}$$

上述权重矩阵的距离规则(即给出两两之间的距离,在此距离内视为两两相邻;否则,不相邻):

$$W_{ij}(d) = \begin{cases} 1 & \text{位置 } j \text{ 与位置 } i \text{ 在距离范围 } d \text{ 之内} \\ & \text{(即位置 } i \text{ 与位置 } j \text{ 相邻);} \\ 0 & \text{位置 } j \text{ 和位置 } i \text{ 在距离范围 } d \text{ 之外} \\ & \text{(即位置 } i \text{ 与位置 } j \text{ 不相邻)。} \end{cases}$$

$$i = 1,2,\cdots,n; j = 1,2,\cdots,m。 \tag{3}$$

上述两个规则可以分别使用,也可以同时使用。如果两两对象客观上空间不相邻,但它们之间在研究的某一方面存在着紧密的联系,应该视为相邻关系,此时即用到距离规则;可见,建立空间权重矩阵距离规则的目的是为了调整确认合理距离内的空间邻接关系。

2.2　方差图[9−11]

方差图的种类很多,有理论方差图、实验方差图、方差云图等,它们从不同的侧面反映了数据在空间上的联系或变化趋势等隐含知识信息。

2.2.1　理论方差图

$$\gamma(h) = \frac{1}{2}E[(z(x) - z(x'))^2] \tag{4}$$

2.2.2　实验方差图

$$\gamma(h_k) = \frac{1}{2|N(h_k)|}\sum_{i=1}^{N_k}[z(x_i) - z(x'_i)]^2 \tag{5}$$

式中:$N(h_k) = \{(i,j): x_i - x_j = h\}$, $|N(h_k)|$ 是 $N(h_k)$ 的不同元素的个数。实验方差图描述了所要研究的数据与距离之间的相关程度,从中可获取不同空间位置的数据之间关系的知识。

2.2.3　方差云图

对 $\forall h$,有 $(z_{i+h} - z_i)^2/2$ 或 $\sqrt{(|z_i + h - z_i|)}/2$ (6)

方差云图可以用来检验潜在的异常值和变化趋势,并且随着距离的增加可以用来估计变化性。通过观察在短距离内产生较大不一致的现象,可以检测出非正常和不均一的地区等有关潜在知识。

2.2.4　方差图的几何异向指(异质方差图)

当空间自相关性随方向发生改变时,就会产生各向异性;此时,方差图既与不同点之间的相互距离有关,又与点的空间配置方向有关。因此,能够反映所研究的对象在不同方向上空间配置的差异性。

另外还有稳健方差图、协方差图、模型方差图以及协相关方差图等。

2.3　空间相关分析技术

为了探索和研究空间对象的空间分布模式(空间依赖和空间异质性),必须利用一些标准全程和局部空间统计,包括 Moran I, Geary C, G 统计和 LISA[12−16]等,它们是 ESDA 技术的基本核心内容之一。

2.3.1　空间自相关检验——Moran I 和 Geary C

(1) Moran I 的定义为:

$$I(d) = \frac{\sum_i^n \sum_{j\neq i}^n w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{S^2 \sum_i^n \sum_{j\neq i}^n w_{ij}} \tag{7}$$

其中,$S^2 = \frac{1}{n}\sum_i^n (x_i - \bar{x})^2$, x_i 表示在 i 处有关对象的观察值。$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$; w_{ij} 是空间权重值,表示为(2)式或(3)式。

利用 Moran I 可以测度空间自相关性,发现观测值在空间分布的差异性和相关性。当位于一定距离 d 内的观察值相近时,Moran I 显著而且为正,不相近时为负,当观察值随机排列且在空间分布的时候则为零[17]。

(2) Geary C 的定义如下:

$$C(d)=\frac{(n-1)\sum_{i}^{n}\sum_{j}^{n}w_{ij}(z_i-z_j)^2}{2nS^2\sum_{i}^{n}\sum_{j}^{n}w_{ij}}$$

$$=\frac{(n-1)}{2n^2}\sum_{i}^{n}C_1(d) \quad (8)$$

Geary 统计总是正的不对称正态分布。Geary 统计的假设检验是用 Geary 统计的平均值，当没有空间自相关性时为1。显著且低的值(在0,1之间)表明了正的空间自相关，而显著且高的值(大于1)表明了空间负相关[14]。

2.3.2 局部空间相关检验——G 统计和 LISA

(1) G 统计

G 统计的定义如下[15]：

$$G_i(d)=\frac{\sum_{j,j\neq i}^{n}w_{ij}x_j}{\sum_{j,j\neq i}^{n}x_j} \quad (9)$$

其中，x_i 表示在 i 处有关对象的观察值，w_{ij}如(3)式。其标准化形式定义为[18]：

$$z(G_i)=\frac{\sum_{j\neq i}^{n}w_{ij}(d)(x_j-\bar{x}_i)}{S_i\sqrt{w_i(n-1-w_i)/(n-2)}} \quad (10)$$

其中 $i\neq j$, $\bar{x}_i=\frac{1}{n-1}\sum_{j,j\neq i}^{n}x_j$, $w_i=\sum_{j,j\neq i}^{n}w_{ij}(d)$, $S_i^2=\frac{1}{n-1}\sum_{j,j\neq i}^{n}(x_j-\bar{x}_i)^2$。显著且为正的 $z(G_i)$表明位置 i 周围的值相对大；反之，显著且为负表明位置 i 周围的值相对小。G 统计可以用来确定空间聚类的模式：高值簇或低值簇。

(2) 空间联系的局部指示(LISA)

Anselin 认为，局部 Moran 和 Geary 统计是可选择的局部指示量[16]。局部 Moran 确定空间聚类模式时，与 G 统计相似；局部 Geary 可以确定相似或不相似的空间模式。局部 Moran 和局部 Geary 的一个好处是它们可以与全程统计(Moran I 和 Geary C)相联系，估计独立统计量对相应的全程统计量的贡献。

局部 Moran I

每一个观察值 i 的局部 Moran 统计量的定义如下[16]：

$$I_i(d)=z_i\sum_{j\neq i}^{n}w_{ij}z_j \quad (11)$$

其中，z_i 和 z_j 的观察值是标准化形式。局部 Moran 的解译与 G 统计相似。I_i 的伪显著水平 p 值可以通过条件随机化或重排的方法衡量[16]，小的 p 值(比如 $p<0.05$)指出位置 i 的周围值相对较高。大的 p 值(比如 $p>0.95$)表明位置 i 周围值相对较低。

局部 Geary

每个观察值 i 的局部 Geary 统计量定义如下[16]：

$$C_i(d)=\sum_{j\neq i}^{n}w_{ij}(z_i-z_j)^2 \quad (12)$$

式中，z_i 和 z_j 是标准化后的值。伪显著水平 p 值与局部 Moran 统计相似。大的 p 值(比如 $p>0.95$)表明极端小的 C_i 值，说明位置 i 的观察值与周围的观察值是正的空间联系(相似)，而小的 p 值(比如 $p<0.05$)表明极端的大的 C_i 值，说明位置 i 的观察值与周围的观察值是负的空间联系(不相似)。

3 应用实例

3.1 ESDA 技术与 GIS 结合的知识发现

3.1.1 S-PLUS 和 ArcView 的结合

GIS 和传统的(空间)统计分析是一种松散耦合的关系，没有形成一种紧密联系的内部机制，至多是一种数据传递和应用的关系，即首先通过统计分析软件进行有关计算，然后把计算结果通过中间软件、程序软件接口转换到 GIS 软件中进行有限的分析应用。这种松散耦合的连接方式限制了目前大部分统计分析软件的进一步拓展，也使得从 GIS 数据库中发现知识的研究受到了一定的限制。美国 MathSoft 公司最近推出的 S-PLUS for ArcView 是 ArcView 的一个扩展模块。它集成了 S-PLUS 4.0 的统计分析功能和 ArcView 的空间可视化技术，是空间统计分析和 GIS 软件结合的一个典范，可用于知识发现，并为其实际应用提供了一个可行的思路(图1)。通过它可在 ArcView 中调用或关闭 S-PLUS，能够从 ArcView 图形文件中选择记录或变量输出到一个 S-PLUS 变量；还能够直接从 ArcView 中得到一个 S-PLUS 对象，如从 ArcView 中的 S-PLUS 模块获取残差或合适的数据来绘图，能够进行各种空间分析，如建立空间权重，计算空间自相关和估计空间模型。

3.1.2 现有 GIS 软件空间统计分析功能扩展的一般方法

GIS 软件与空间分析软件结合的关键在于数据的读取和存储，在此基础上进行数据分析和结果表达。二者结合的主要方式如图2，结合的主要目的是在 GIS 软件中增加统计分析功能，为统计分析结果提供先进的可视化表达方法，便于从 GIS 数据库

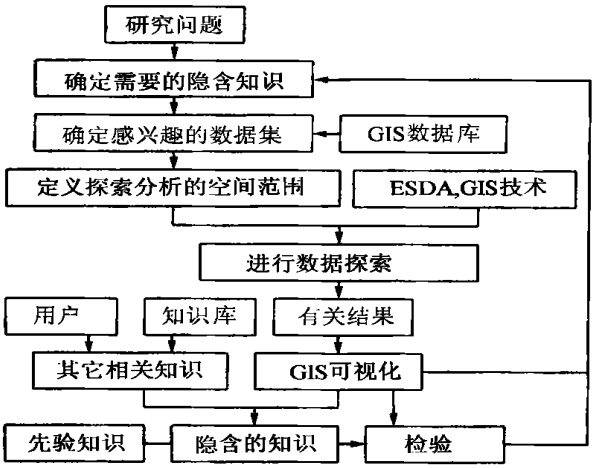

图 1　用 ESDA 技术从 GIS 数据库中发现知识的基本框架

Fig. 1　Basic flwo chat of knowledge discovery with ESDA from GIS database

中发现隐藏的知识规律。

3.2　中国流动人口空间分布规律的知识发现

利用 1996 年中国资源环境数据库(1∶400 万)[18]中行政单元数据层(图形数据,以县为单位)和中国公安部 1996 年流动人口统计数据(属性数据,以县为单位)作为原始数据,通过 ArcView 软件、S-PLUS for ArcView 模块、S-PLUS 中的 PLUS+语言以及 ArcView 中的 Avenue 语言进行编程,从中发现中国流动人口空间分布规律的有关知识。

(1) 观察、调整中国流动人口的方差云图发现广东深圳地区的流动人口明显异常。

(2) 对流动人口进行全程自相关分析(计算 Moran I)发现,$p=1.109\times10^{-6}$正态统计量为 4.910,远大于正态分布函数在 0.05 水平下的数值

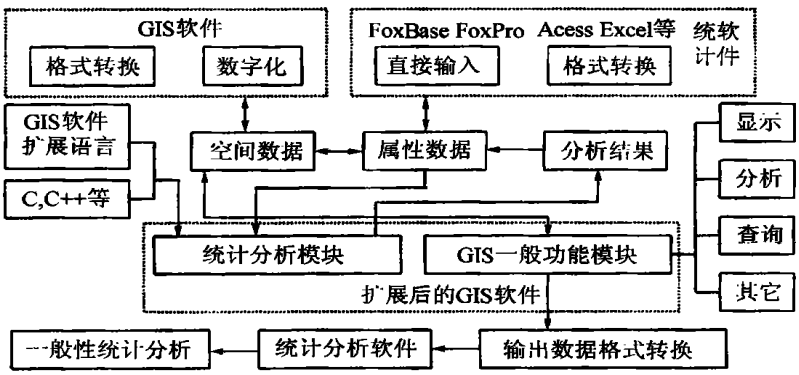

图 2　GIS 与空间分析的紧密结合

Fig. 2　GIS's integrating compactly with spatial analysis

(1.645)。这表明中国流动人口在空间上具有显著的相关性,不是随机分布的,具有一定的规律,存在必然的内在联系。进行更深入的分析可发现深层次的隐含知识。

(3) 中国的流动人口主要集中在城市,为了初步揭示流动人口的空间差异,我们选取了流动人口在 10 万以上的 84 个城市利用异质方差图(图 3)进行空间差异性分析。结果表明,北京方向、深(圳)广(州)方向、上海方向的不同距离的方差变化范围最大,说明这 3 个空间方向的城市流动人口偏离均值的范围最大,联系这 3 个方向的实际地理位置即可判断这 3 个方向的城市流动人口很多,与其它地区的差异性很大。

(4) 为了进一步揭示中国流动人口的分布规律,我们对全国每个县的流动人口进行局部自相关分析(计算局部 Moran I 值),局部 Moran I 值的空间分布如图版 I 图 4。图版 I 图 4 中同一种颜色地区内部的流动人口相似性大,不同颜色地区之间差异性大。

(5) 单从图版 I 图 4 无法准确分析中国流动人口的空间分布规律,需要一些其它知识如数字人口高程模型以及由此产生的等值线(图版 I 图 5),然后把它们与行政单元界线相叠加,采用定性和定量相结合的方法发现规律知识。

(6) 通过分析,结合已有的地理知识,可发现如下隐含知识:

• 局部 Moran I 值小于 0 地区的面积占全国的 6.3%,流动人口占全国的 33.4%。这类地区绝大部分为城市市区及其周围地区,是流动人口聚集区。其内部差异性很大,与外部的差异性也很大。

• 局部 Moran I 值 0—0.2 地区的面积占全国的 18.5%,流动人口占全国的 51.7%。这类地区也是中国流动人口的聚集区,流动人口数量较多,但内部的相似性较小。

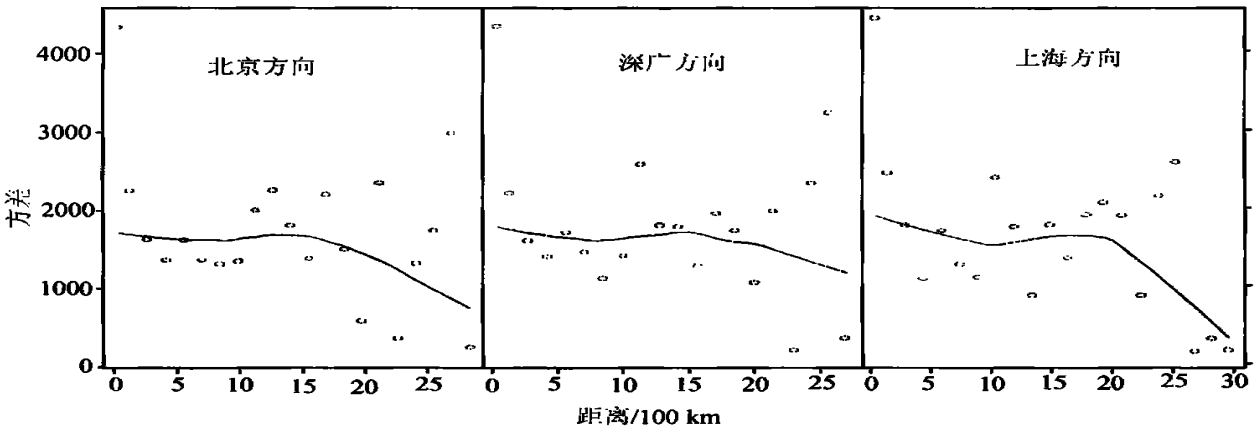

图 3 不同方向的中国城市流动人口异质方差图

Fig. 3 Anisotropic variogram of the floating population in different direction in Chinese cities

• 局部 Moran I 值为 0.2—0.8 地区的面积占全国的 75.1%，流动人口占全国的 18.3%，这类地区流动人口数量很少，但内部的相似性很大。

• 局部 Moran I 值大于 0.8 地区的面积占全国的 0.2%，流动人口占全国的 14.9%。这类地区是中国流动人口的核心区，主要集中在珠江三角洲、长江三角洲和京津地区，这类地区不仅流动人口数量大，而且内部的相似性也大。

(7) 对上述发现的知识进行整理加工，还可按照全国流动人口状况的内部相似性和外部差异性并结合数字人口模型进行分区分类研究。

4 结论与讨论

(1) 从 GIS 数据库中发现知识需要综合知识的支持与结合，学科交叉、知识综合等在这一领域体现的最为充分；靠一个学科、一门技术难以获得客观有效的实用性知识，至多获得一些初步的、肤浅的知识且难以进入实用阶段；先验知识以及相关知识往往起到深层次发现知识的基石作用。

(2) 通过 ESDA 技术从 GIS 数据库中探索空间分布规律是可行的、有效的，但必须充分利用 GIS 的可视化技术和空间分析技术，实现 ESDA 与 GIS 的紧密完全结合，特别是二者的软件集成。

(3) 完全自动地从 GIS 数据库中发现知识是不可能的或者说发现的知识是有限的；本文所描述的从 GIS 数据库中发现知识的方法是人机交互的，实现这一过程，需要良好的人机交互界面以及 GIS 可视化的支持；因此，ESDA 技术的分析结果如何有效方便地通过 GIS 可视化方法表达出来，是我们应该重点研究的问题。

(4) 从 GIS 数据库中发现知识需要专家系统的支持，最终实现它们的完全集成，实现智能化 GIS，直接为辅助决策、制定政策法规服务。

参 考 文 献 (References)

[1] Li D R, Cheng T. Knowledge Discovery from GIS Databases [J]. *Acta Geodaetica et Cartographica Sinica* 1995, **24**(1): 37—44. [李德仁，程涛. 从GIS 数据库中发现知识[J]. 测绘学报，1995，**24**(1)：37—44.]

[2] Di K C, Li D R *et al*. A Framework of Spatial Data Mining and Knowledge Discovery [J]. *Journal of Wuhan Techinical University of Surveying and Mapping*, 1997, **24**(1): 6—10. [邸凯昌，李德仁等. 空间数据发掘和知识发现的框架[J]. 武汉测绘科技大学学报，1997，**22**(4)：328—332.]

[3] Di K C, Li D R *et al*. Rough Set Teory and Its Application in Attribute Analysis and Knowledge Discovery in GIS [J]. *Journal of Wuhan Techinical University of Surveying and Mapping*, 1999, **24**(1): 6—10. [邸凯昌，李德仁等. Rough 集理论及其在 GIS 属性分析和知识发现中的应用[J]. 武汉测绘科技大学学报，1999，**24**(1)：6—10.]

[4] Gao J. Visualization in Geo-Spatial Data[J]. *Engineering of Surveying and Mapping*, 2000, **9**(3): 1—7. [高俊. 地理空间数据的可视化[J]. 测绘工程，2000，**9**(3)：1—7.]

[5] Liu M L, Li Q. From Quantitative Geography to Geocomputation: Reconsidering Geographical Quantitative Methods [J]. Human Geography, 2000, **5**(3): 13—16. [刘妙龙，李乔. 从数量地理学到地理计量学——对数量地理方法的若干思考[J]. 人文地理，2000，**5**(3)：13—16.]

[6] Haslett J, Wills G *et al*. SPIDER-An Interactive Statisticcal Tool for the Analysis of Spatially Distributed data [J]. *I. J. Geographical Information System*, 1990, **4**(3): 285—296.

[7] Batty M, Xie Y. Modelling inside GIS: Part 1: Model Structures, Exploratory Spatial Data Analysis and Aggregation [J]. *I. J. Geographical Information System*, 1994, **8**(3): 291—307.

[8] Bao Y C, Li X *et al*. Spatial Data Analysis and Spatial Models[J]. *Geographical Reseach*, 1999, **18**(2): 185—190. [柏延臣,李新等. 空间数据分析和空间模型[J]. 地理研究,1999,**18**(2): 185—190.]

[9] Matheron G. Principles of Geostatistics [J]. *Economic Geology*, 1963, **58**(2): 1246—1266.

[10] Cressie Noel A, Hawkins D M. Robust Estimation of the Variogram [J]. *Journal of the International Association for Mathematical Geology*, 1980, (12): 115—125.

[11] Cressie Noel A. Statistics for Spatial Data [M]. NY:John Wiley & Sons, Inc. 1993.

[12] Chen Y, Chen W W. Data Mining and Statistics [J]. *Computer Engineering & Application*, 2000, (5): 15—17. [陈元,陈文伟. 数据开采与统计学. 计算机工程与应用,2000,(5):15—17.]

[13] Cliff A D, Ord J K. Spatial Autocorrelation [M]. Poin, London, 1973.

[14] Cliff A D, Ord J K. Spatial Processes: Models and Applications [M]. Pion, London, 1981.

[15] Getis A., Ord J K. The Analysis of Spatial Association by the Use of Distance Statistics [J]. *Geographical Analysis*, 1992 (24): 189—206.

[16] Anselin L. Local Indicators of Spatial Association:LISA [J]. *Geographical Analysis*, 1995, **27**(3): 93—115.

[17] Goodchild M F, Haining R P, Wise S. Integrating GIS and Spatial Data Analysis:Problems and Possibilities [J]. *International Journal of Geographical Information Systems*, 1992, **6**(5): 407—423.

[18] State Key Laboratory of Resources and Environment of the Institute of Chinese Academy. Database for Chinese Resources and Environment (1:4000000) [R]. 1996. [中国科学院资源与环境信息系统国家重点实验室. 中国资源环境数据库(1:400 万)R. 1996.]

Knowledge Discovery with ESDA from GIS Database

MA Rong-hua, HUANG Xing-yuan, ZHU Chuan-geng

(*Institute of GIS & RS, Nanjing University, Nanjing* 210093, *China*)

Abstract: Lots of knowledge can be discovered from GIS database. Knowledge discovery from database of GIS has been drawing more and more attentions in different fields such as geo-science field, and computer field. ESDA technique especially emphasizes the study on the spatial dependency and spatial heterogeneity. And it is used to preliminary discover some law hidden in the data sub-aggregate of interest in the field of knowledge discovery. It is one of the most important contents to some normal whole and local spatial statistics including Moran I, Geary C, G statistics, LISA and so on. And it is the key of making ESDA technique and GIS completely and closely integrated to fully make use of GIS visualization and spatial analysis technique under the condition of imperfect Expert System. The example, which is China floating population's spatial distribution law discovered from GIS database, shows the method and steps to discover knowledge from database of GIS with the integrating ESDA and other knowledge. The results show that it is practical and effective to discover knowledge from from GIS database with ESDA. But it is impossible to completely and automatically discover knowledge from GIS database and other synthetic knowledge must be integrated into the course of knowledge discovery. The method of knowledge discovery in the paper is used in the mode of man-machine conversation and it needs urgently to be supported by Expert System.

Key words: ESDA; GIS; Database; Knowledge

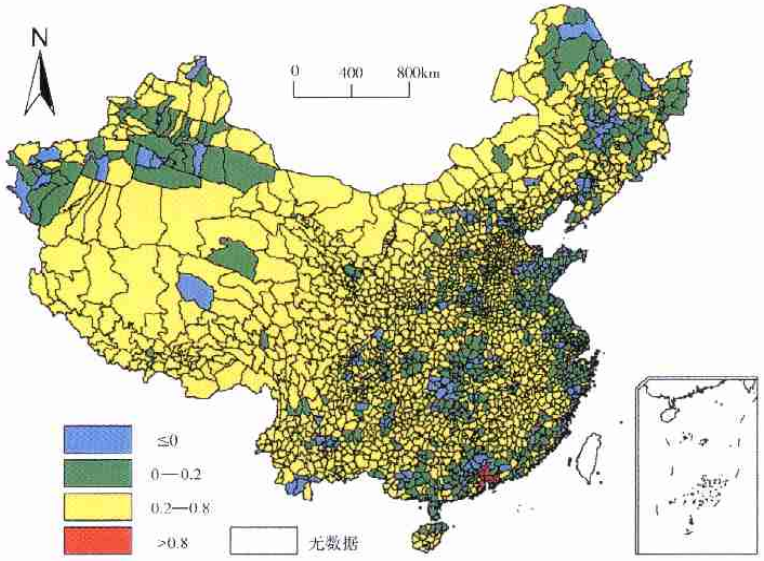

图 4 局部 Moran I值的空间分布（以县/市为单位，中国台湾地区无流动人口数据）

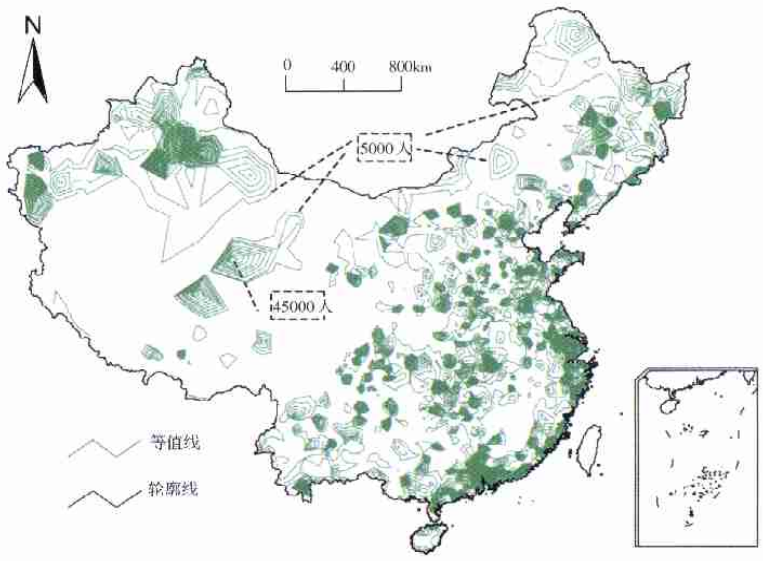

图 5 1996 年中国流动人口等值线图（等高距为 5000 人，中国台湾地区无流动人口数据）