

GIS 的不同次区域之间相似性度量

廖伟华

(广西大学数学与信息科学学院, 南宁 530004)

摘要: GIS 属性相似度量有相似系数及以空间位置的空间自相关。这 2 种方法都不能计算全域为背景的次区域相似性问题, 而以隶属函数的粗糙度量很好地解决了这个问题。本文利用粗糙集计算了 GIS 的离散型数据的次区域相似度量。利用邻域粗糙集计算了连续属性值的上下近似逼近, 并利用邻域信息粒子计算连续属性值的粗糙隶属问题, 从而计算连续属性值次区域相似度量问题。本文方法能度量次区域而不是全域或单个元素之间的空间相关与自相关问题, 考虑了全区域背景下的 GIS 相似问题。并针对 GIS 连续属性值, 利用距离函数和邻域粗糙集来划分连续属性的上下近似, 以及分类问题, 提出一种基于邻域信息粒子的粗糙隶属函数。最后利用粗糙相似度量公式度量 GIS 次区域的相似问题。

关键词: 次区域; 粗糙集; 邻域粗糙; 相似度量

DOI: 10.3724/SP.J.1047.2012.00426

1 引言

近年来, 相关性研究在人口、地学、经济等领域广泛开展。传统意义的相关性度量指标一般为相关系数。而 GIS 研究由于具有空间概念, 因此, 相关性研究相对于其他学科更有空间的含义。一般在 GIS 相关性度量研究中, 有相关系数与空间自相关。近年来, 国内外对于相关性的研究一般都集中在其应用上, 如人口、DEM 和气温等方面的连续数据相关性的数据分析上^[1-8]。在 GIS 背景下, 相关系数要求两个相等序列的连续数据, 而自相关则是研究空间内部的全局与局部相关性。对于一个区域的两个次区域之间的离散型分类数据, 如何在考虑全局背景下去度量两个次区域的相似性, 这 2 种方法就无能为力了。因此, 本文利用粗糙次空间, 以及粗糙隶属函数度量两个不同次区域的相似性问题。

2 空间自相关与相关系数

2.1 全局空间自相关

全局空间自相关是通过对属性值在一整个区域内与其他的空间目标之间特征的描述, 主要的测

量指标有 Global Moran's I 和 Global Geary's C 等, 它主要分析区域总体的空间差异程度和空间关联。最常用的指标是 Moran's I, 其公式见式 1:

$$I = \frac{n}{S_0} \cdot \frac{\sum_{i=1}^n \sum_{j=1}^n \omega_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (1)$$

其中, x_i 是第 i 个样本的属性值, \bar{x} 是 x_i 样本的平均值, S_0 是所有空间对象之间位置权重矩阵之和。

$$S_0 = \sum_{i=1}^n \sum_{j=1}^n \omega_{ij} \quad (2)$$

式中, ω_{ij} 为第 i 与 j 个空间目标对象之间的连接矩阵, 即 n 阶矩阵, 一般情况下可以通过空间对象之间的拓扑属性, 如空间邻接性来建立, 也可以通过空间对象之间的距离来建立, 如果两个对象之间的距离小于指定阈值, 则 $\omega_{ij} = 1$, 其他情况等于 0。Moran's I 计算完成之后, 一般可以通过 z 检验统计计算结果:

$$Z(I) = \frac{I - E(I)}{\sqrt{VAR(I)}} \quad (3)$$

Moran's I 是一个空间自相关系数, 其值为 $(-1, 1)$ 。在指定检验置信性水平下, 当其值显著为正时, 表示观测样本对象之间是一种显著的正相关关系, 低属性值的空间对象和低属性值对象聚在一

收稿日期: 2011-07-11; 修回日期: 2012-07-21.

基金项目: 广西自然科学基金项目(2010GXNSFA013109); 广西大学科研基金资助项目(XJZ110584)。

作者简介: 廖伟华(1975-), 汉族, 男, 广西大学讲师, 主要从事 GIS 不确定性研究。E-mail: gisliao@163.com

起,高属性值的对象和高属性值集对象聚在一起,为低低集聚或高高集聚;当其值显著为负时,表示空间对象之间观测值是一种显著的负相关关系,低属性值空间对象和高属性值空间对象聚在一起,是一种分散空间格局;当其值趋于 0 时,表明空间对象之没有空间自相关关系,观测值在空间位置上是随机格局^[9]。

20	30	40
50	47	24
23	22	10

图 1 自相关属性取值分布图
Fig. 1 Autocorrelation value map

1	2	1
2	2	3
1	2	2

图 2 局部自相关分布图
Fig. 2 Local autocorrelation map

设有如图 1 的 9 个面域多边形,编号从左至右,从上往下依次为{1,2,3,...9},图中标识为每个多边形的属性值,其中取值为随机产生的连续空间取值。这 9 个多边形全局 moran's I 值为 0.028508,Z 值为 0.636673。可以看出图 1 的分布是一种分散型随机排列的空间格局。

2.2 局部空间自相关

全局空间自相关统计量是一种空间自相关总

体统计指标,它只说明所有空间对象与周边空间对象之间空间差异的平均差异。有时,当 Moran's I 值缩小时,空间对象局部空间差异有可能是在扩大。ESDA 局部分分析方法是一种反映区域经济空间差异的变化趋势分析方法。Anselin 在 1994 年提出了空间对象联系的局部指标 LISA(Local Indicators of Spatial Association),它揭示的空间自相关性质为局部到每个空间单元对象之间^[10]的关系。LISA 其实是将 Moran's I 分解到各个统计空间单元^[11],每一个统计单元 i 的统计值可表示为:

$$I_i = \sum \omega_{ij} z_i z_j \quad (4)$$

式中, Z_i 与 Z_j 是不同空间对象属性值的标准化平均值, ω_{ij} 是空间对象权重连接矩阵。这样在一定统计置信水平下,若 I_i 大于 0,且 z_i 小于 0,则表明空间对象 i 和周围空间对象的观测值都相对较低,是低低集聚;若 I_i 大于 0,且 z_i 大于 0,则表明空间对象 i 和其周围空间对象的属性值都相对较高,是高高集聚;若 I_i 小于 0,且 z_i 小于 0,则表明空间对象 i 的观测值远小于其周围空间对象的观测值,是低高集聚。若 I_i 小于 0,且 z_i 大于 0,则表明空间对象 i 的观测值远大于其周围空间对象的观测值,是高低集聚;

同样对图 1 进行局部 moran's I 运算,可以得到图 2 的局部空间自相关分布图(图 2)。其中,1 代表低高集聚,2 代表高高集聚,3 代表高低集聚。

由上可知,空间自相关的计算有以下性质:

(1)只能针对连续性的属性值进行求解,对于离散型的分类数据无能为力。

(2)自相关只能计算整体或是每个单元元素的相关性问题,但对于几个单元组成的次区域与其他次区域的相关性,则不能计算。

2.3 相关系数

统计中常用相关系数 r 来衡量 2 个变量之间的线性相关的强弱,当 x_i 不全为零, y_i 也不全为零时,则 2 个变量的相关系数的计算公式是:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (5)$$

式中, r 为变量 y 与 x 的相关系数(简称相关系数)。其中, \bar{x}, \bar{y} 分别为序列 x_i 与 y_i 的平均值。很显然我们可以得到以下的一些性质

(1) 相关系数公式只能针对连续性的属性值进行的求解, 对于离散型的分类数据无能为力;

(2) x_i 与 y_i 两个序列必须长度相同, 长度不同就不能进行求解。

3 次区域研究的粗糙度量相关概念

定义 1: 设 U 为论域, R 是 U 上的等价关系, $\forall A \subseteq U$, 元素 $x \in U$ 在集合 A 中的粗隶属度为^[12]:

$$\mu_A^R(x) = \frac{|A \cap [x]_R|}{|[x]_R|} \quad (6)$$

即 x 为 A 中的粗隶属度是 x 所处的等价类 $[x]_R$ 在模糊集“弱包含于 A ”中的模糊隶属度; 可以将粗隶属度理解为一个系数, 它描述了一个 $x \in U$ 是 A 的成员的精确性。

性质 1: 设 U 为论域, R 为 U 上的等价关系, $\forall A \subseteq U$, 则由 R 和 A 能确定 U 上的模糊集^[13]:

$$\begin{aligned} \mu_A^R: U &\rightarrow [0, 1] \\ x &\rightarrow \mu_A^R(x) \end{aligned} \quad (7)$$

显然, 上述公式都是建立在离散数值空间的分明等价关系之上, 等价关系对论域的划分形成了论域空间的粒化^[14]。然而对于实数空间而言, 对象的取值是连续的, 如 DEM 的属性值等。对于此类空间, 采用等价关系将导致对个别数值属性的过拟合。邻域结构和序结构是实数空间的重要结构, 对于连续数值, 即给出其邻域结构。

一般可以对邻域进行 2 种方法定义: 一是计算对象邻域内对象的数量, 如经典的 k -近邻方法; 另一种是计算对象边界到度量邻域中心点的最大距离。本研究中, 我们采取第 2 种方法计算邻域。

定义 2: 给定实数空间 Ω , 其为 N 维的, d 是 R^N 上的一个度量, 如果 d 满足^[14]:

(1) $d(x_1, x_2) \geq 0, d(x_1, x_2) = 0$ 当且仅当 $x_1 = x_2, \forall x_1, x_2 \in R^N$;

(2) $d(x_1, x_2) = d(x_2, x_1), \forall x_1, x_2 \in R^N$

(3) $d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3), \forall x_1, x_2, x_3 \in R^N$

这样, (Ω, d) 是一个度量空间。我们常用欧氏距离进行实数空间上常用的距离度量。

定义 3: 设有一实数空间 $U \{x_1, x_2, x_3, \dots, x_n\}$, x_i 是 U 上的任意对象, 其 δ -邻域定义为^[14]:

$$\delta(x_i) = \{x \mid x \in U, d(x, x_i) \leq \delta\} \quad (8)$$

$\delta(x_i)$ 是一个 δ 邻域信息粒子, 它由描述对象 x_i

生成, 简称为 x_i 的邻域粒子, 其中 $\delta > 0$ 。

而定义 1 基于 pawlak 粗糙集等价类的粗糙隶属度, 可当一个 GIS 属性为连续值时, 不能随便将其划分为等价类。针对连续值, 我们按照邻域结构, 给出邻域粗糙隶属度的定义。

定义 4: Ω 是一个 N 维的实数空间, 其中, U 为论域, $\delta(x_i)$ 为对象 x_i 的邻域粒子, $\forall A \subseteq U$, 元素 $x \in U$ 在集合 A 中的粗隶属度为^[15]:

$$\mu_A^R(x) = \frac{|A \cap \delta(x_i)|}{|\delta(x_i)|} \quad (9)$$

即 x 所在的邻域信息粒子在模糊集“弱包含于 A ”中的模糊隶属度是 x 在 A 中的粗隶属度。

定义 5: 给定论域 $U = \{u_1, u_2, \dots, u_n\}$, R 为 U 上的等价关系, A 与 B 是 U 上的两个不同的粗糙集, $A, B \subseteq U, u_i \in U, u_i$ 在等价关系 R 下关于 A 与 B 的粗隶属度分别是 $a_i = \mu_A^R(u_i)$ 和 $b_i = \mu_B^R(u_i) (i = 1, 2, \dots, n)$, 当在等价关系 R 下, 则 A 和 B 的粗隶属函数分别表示为 A', B' ^[13]:

$$\begin{aligned} A' &= \frac{\mu_A^R(u_1)}{u_1} + \frac{\mu_A^R(u_2)}{u_2} + \dots + \frac{\mu_A^R(u_n)}{u_n} \\ B' &= \frac{\mu_B^R(u_1)}{u_1} + \frac{\mu_B^R(u_2)}{u_2} + \dots + \frac{\mu_B^R(u_n)}{u_n} \end{aligned} \quad (10)$$

则下列公式可以计算集合 A 和 B 之间的相似程度^[13]:

$$SimD_R(A, B) = \begin{cases} 1, & A = B = \Phi \\ \frac{\sum_{i=1}^n \min \{a_i, b_i\}}{\sum_{i=1}^n \max \{a_i, b_i\}} & else \end{cases} \quad (11)$$

本文利用文献[15]给出的 A, B 相似度的计算公式^[15]:

$$SimD_R(A, B) = \begin{cases} 1, & A = B = \Psi \\ \frac{2 \sum_{i=1}^n \min \{a_i, b_i\}}{\sum_{i=1}^n (a_i + b_i)} & else \end{cases} \quad (12)$$

显然, 当集合 A 与 B 之间的相似程度越大, 则 $SimD_R(A, B)$ 会越大; 反之亦然。且公式满足下列性质:

(1) $SimD_R(A, B) \in [0, 1]$;

(2) $SimD_R(A, B) = SimD_R(B, A)$;

(3) $SimD_R(A, B) = 0$, 当仅对 $\forall u_i \in U (i=1, 2, \dots, n)$, $\mu_A^R(u_i)$ 和 $\mu_B^R(u_i)$ 至少有一个为 0, 且集合 A 和 B 不能同时为空集。

因此, 对于 GIS 次区域研究, 我们比较两个次区域的相似性, 并不需要确定两个次区域的形状和结构。对于离散型变量, 按照离散粗隶属度公式, 我们只需确定该单元所在等价类与次区域的交集个数。对于连续型变量, 按照连续型粗隶属度公式, 我们只需确定该单元的邻域与次区域的交集个数。

4 GIS 次区域间相似性度量的实例分析

设有如图 3 的 100 个栅格多边形。编号从上至下, 从左至右分别为 $\{x_1, x_2, \dots, x_{100}\}$, 值域为 $\{1, 3, 4\}$ 。今有如图 3 中 A, B, C 3 个次区域覆盖多边形, 每个次区域都覆盖 16 个单元栅格多边形, 如何度量这 3 个次区域多边形之间的相似性呢? 我们以 x_1 为例, 它所在的等价类为离散值 1 的分类, 涉及区域分别为 $\{x_1, x_2, x_3, x_4, x_5, \dots, x_{94}\}$ 共 58 个, 而 x_1 所在的等价类在次区域 A 中有 $\{x_{11}, x_{12}, \dots, x_{53}\}$ 共 10 个。则 x_1 的粗隶属度为 $10/58$, 结果为 0.17。由粗隶属函数的公式计算有:

$$A = \frac{0.17}{x_1} + \frac{0.17}{x_2} + \frac{0.17}{x_3} + \frac{0.16}{x_4} + \frac{0.17}{x_5} + \frac{0.16}{x_6} + \dots + \frac{0.16}{x_{100}}$$

$$B = \frac{0.12}{x_1} + \frac{0.12}{x_2} + \frac{0.12}{x_3} + \frac{0.22}{x_4} + \frac{0.12}{x_5} + \frac{0.12}{x_6} + \dots + \frac{0.22}{x_{100}}$$

$$C = \frac{0.16}{x_1} + \frac{0.16}{x_2} + \frac{0.16}{x_3} + \frac{0.16}{x_4} + \frac{0.16}{x_5} + \frac{0.16}{x_6} + \dots + \frac{0.16}{x_{100}}$$

则次区域 A 和 B 之间的相似度量为:

$$SimD_R(A, B) = \frac{2 \sum_{i=1}^{100} \min \{a_i, b_i\}}{\sum_{i=1}^{100} (a_i + b_i)} = \frac{2(0.12 + 0.12 + 0.12 + 0.16 + 0.12 + \dots + 0.16)}{(0.12 + 0.17) + (0.12 + 0.17) + \dots + (0.16 + 0.22)} = 0.8080$$

同理: $SimD_R(A, C) = 0.9515$, $SimD_R(B, C) = 0.8594$ 。因此, A 与 B 之间的相似性小于 A 与 C 之

间的相似性, A 与 C 之间的相似性大于 B 与 C 之间的相似性。

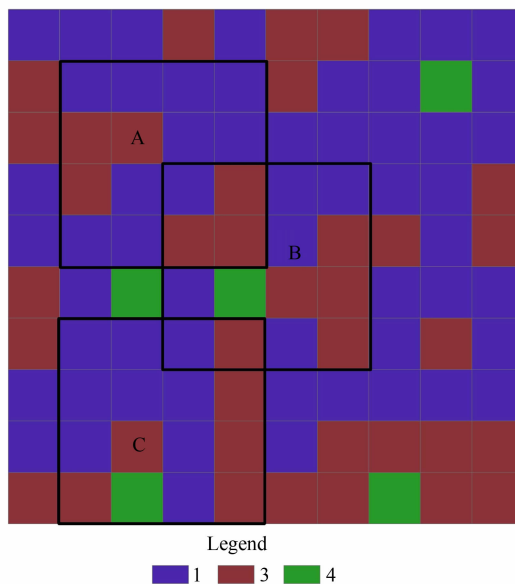


图 3 全域多边形分类及次区域分布(离散数据)

Fig. 3 Universe polygon classification and subzone map (discrete data)

同样对于图 4 的连续取值, 由于属性为一个, 我们采用绝对距离公式, $\delta=10$ 对图进行邻域粒化。依次可以得到每个多边形元素之间的距离, 以及每个多边形元素的邻域粒子。如第一个多边形的邻域粒子为 $\{1, 10, 27, 28, 34, 50, 51, 65, 68, 75, 94, 98, 99, 100\}$, 在区域 A 中的粗糙隶属度为 $1/14$, 在区域 B 中的粗糙隶属度为 $2/14$, 在区域 C 中的粗糙隶属度为 $2/14$ 。按照粗隶属函数的公式计算有:

$$A = \frac{0.07}{x_1} + \frac{0.07}{x_2} + \frac{0}{x_3} + \frac{0.21}{x_4} + \frac{0.29}{x_5} + \frac{0.36}{x_6} + \dots + \frac{0.07}{x_{100}}$$

$$B = \frac{0.14}{x_1} + \frac{0.14}{x_2} + \frac{0.29}{x_3} + \frac{0.14}{x_4} + \frac{0.14}{x_5} + \frac{0.21}{x_6} + \dots + \frac{0.14}{x_{100}}$$

$$C = \frac{0.21}{x_1} + \frac{0.07}{x_2} + \frac{0.07}{x_3} + \frac{0}{x_4} + \frac{0.14}{x_5} + \frac{0.14}{x_6} + \dots + \frac{0.21}{x_{100}}$$

则次区域 A 和 B 之间的相似度量为:

$$SimD_R(A, B) = \frac{2 \sum_{i=1}^{100} \min \{a_i, b_i\}}{\sum_{i=1}^{100} (a_i + b_i)} =$$

$$\frac{2(0.07+0.07+0+0.14+0.14+\cdots 0.07)}{(0.07+0.14)+(0.07+0.14)+\cdots(0.07+0.14)} = 0.9893。$$

同理, $SimD_R(A, C) = 0.9567$, $SimD_R(B, C) = 0.9271$ 。因此, A 与 B 之间的相似性大于 A 与 C 之间的相似性, A 与 C 之间的相似性大于 B 与 C 之间的相似性。A、B、C 由于考虑了全域背景, 因此, 它们之间的比较不再是单纯的考虑自己的属性值, 而是完全考虑了自己所在的大区域背景下与其他单元属性关系, 也就是让它自己所在的等价类或邻域都参与了计算。对于粗糙相似性的结算结果, 不管是离散型的还是连续型的, 两个次区域之间相似度量值越大, 证明这两个次区域在属性取值上越相似, 属性分布格局越相同。

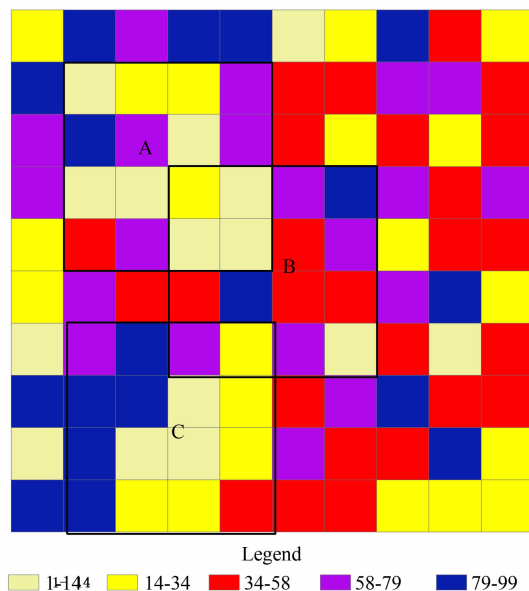


图4 全域多边形分类及次区域分布(连续数据)

Fig. 4 Universe polygon classification and subzone map (continuous data)

5 结论

本文利用粗糙隶属函数度量不同次区域之间的相似性问题。相对于 moran's I, 此方法能度量次区域而不是全域或单个元素之间的空间相关问题, 而是考虑了全区域背景下的 GIS 相似问题。并针对 GIS 连续属性值, 利用距离函数和邻域粗糙集来划分连续属性的上下近似, 以及分类问题, 提出一种以邻域信息粒子的粗糙隶属函数。最后利用

粗糙相似度量公式度量 GIS 次区域的相似问题, 该方法对于 GIS 点群等此目标群的相似度量提供了一种全新的方法。以后的研究将集中在不同分布模式的目标群之间的相似, 以及粗糙熵之间的相似性等问题上。

参考文献:

- [1] 李同升, 王霞. 陕西省非农人口分布的空间自相关特征分析[J]. 西北大学学报(自然科学版), 2007, 37(6): 935-939.
- [2] 宋琳, 董春, 胡晶, 等. 基于空间统计分析与 GIS 的人均 GDP 空间分布模式研究[J]. 测绘科学, 2006, 31(4): 13-125.
- [3] 王劲峰, 李连发, 葛咏, 等. 地理信息空间分析的理论体系探讨[J]. 地理学报, 2000, 55(1): 922-1003.
- [4] 杨凤海, 郭欣欣, 高凤杰, 等. 基于 DEM 聚焦分析的旬平均气温与地面高程的相关性定量研究[J]. 地理与地理信息科学, 2009, 25(6): 37-40.
- [5] 连健, 李小娟, 宫辉力. GIS 支持下的空间分层抽样方法研究——以北京市人均农业总产值抽样调查为例[J]. 地理与地理信息科学, 2008, 24(6): 30-34.
- [6] 程涛, 邓敏, 李志林. 空间目标不确定性的表达方法及其在 GIS 中的应用分析[J]. 武汉大学学报·信息科学版, 2007, 32(5): 389-393.
- [7] 柴思跃, 苏奋振, 周成虎. 基于周期表的时空关联规则挖掘方法与实验[J]. 地球信息科学学报, 2011, 13(4): 455-464.
- [8] 廖伟华. 变精度粗糙集下的 GIS 面目标拓扑关系扩展研究[J]. 地球信息科学学报, 2010, 12(6): 806-810.
- [9] 王耀革, 王志伟, 朱长青. DEM 误差的空间自相关特征分析[J]. 武汉大学学报. 信息科学版, 2008, 33(12): 1259-1262.
- [10] Anselin L. Local indicators of spatial association LISA [J]. Geographical Analysis, 1995, 27: 93-115.
- [11] 张松林, 张昆. 全局空间自相关 Moran 指数和 G 系数对比研究[J]. 中山大学学报(自然科学版), 2007, 46(4): 93-97.
- [12] 刘富春. 模糊粗糙集的相似度量和相似性方向[J]. 计算机工程与应用, 2005, 35: 63-66.
- [13] 王洪凯, 管延勇, 史开泉. 粗集间的相似度量及其应用[J]. 计算机工程与应用, 2004, 31: 39-40.
- [14] 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报, 2008, 19(3): 640-649.
- [15] 史战红, 连玉平. 基于包含度的粗糙集间的相似性度量[J]. 数学教学研究, 2008, 2: 53-54.

The Similarity Measurement for Different Subzones of GIS

LIAO Weihua

(School of Mathematics and Information, Guangxi University, Nanning 530004, China)

Abstract: There are two methods for GIS similarity measurement problems, one is cross-coefficient for GIS attribute similarity measurement, and the other is spatial autocorrelation that is based on spatial location. Both of these two methods can not measure subzone similarity of GIS subzone based on universal background. The rough measurement based on membership function solved this problem well. In this paper we used rough sets to calculate the GIS subzone discrete data similarity measurement, and used neighborhood rough sets to calculate continuous data's upper and lower approximation. We used neighborhood particle to calculate continuous attribute's rough membership function, then to calculate continuous attribute's subzone similarity measurement problem. This paper used rough membership to measure similarity problem for different subzones. Because Moran's I can only measure universe or each unit's spatial autocorrelation, it can not measure subzone, so our method in this paper can compute GIS subzone similarity based on universe. And for continuous value, we used distance function and neighborhood rough sets to divide continuous value's upper and lower approximation and classification problem, then we put forward a rough membership function based on neighborhood information granulation. Then, we used rough similarity measurement formula to measure GIS subzone similarity problem. This method can provide a new direction for GIS point group or others' object group similarity measurement. At last, using an example that includes discrete and continuous value, we can find that spatial autocorrelation can not measure discrete value, cross-coefficient can not measure discrete value too. If the subzone in map is not equal length for continuous value, cross-coefficient can not measure similarity, but the rough measurement based on membership function solved this problem well.

Key words: subzone; rough sets; neighborhood rough sets; similarity measurement