

多边形统计数据空间分析的不确定性研究

——以北京市海淀区人口普查数据为例

张小虎^{1,2}, 钟耳顺¹, 王少华^{1,2}, 张珣^{1,2}, 张济³

(1. 中国科学院地理科学与资源研究所, 北京 100101; 2. 中国科学院大学, 北京 100049;
3. 国家林业局林产工业规划设计院, 北京 100714)

摘要: 普查数据是地理学空间分析的重要数据源。由于受到数据与计算机处理能力的限制, 以往的研究对普查数据空间分析的不确定性未给予足够重视, 也未形成成熟的研究方法。在建筑物单元的人口普查数据支持下, 本文基于多边形统计数据的可塑面积单元问题(Modifiable areal unit problem, MAUP)特征, 设计了一种该类数据空间分析不确定性的研究方法, 采用不同的尺度(Scale)及分区(Zoning)系统对多边形的统计数据空间分析的准确性进行了分析。实验引入尺度与形态指数, 利用可视化分析和数据拟合的研究方法, 对尺度及分区对空间分析结果的影响模式进行了模拟。研究结果表明: (1)以统计小区的空间分析, 其结果受统计小区空间形态的影响较大, 不确定性强, 不能充分反映统计数据本身的空间特征; (2)规则格网能较好地保持原始统计数据的空间分布特征, 但仍然受尺度及分区影响; (3)规则格网的空间分析结果及其准确性与尺度有较好的拟合关系, 不同尺度下的分析结果不确定性是原始数据不同尺度特征的体现; (4)分区效应受空间分析方法的计算尺度影响, 两者共同对空间分析结果产生影响。对于固定尺度的规则格网, 其邻接多边形数目是分析结果不确定的主要原因。本文研究结果表明, 在多边形统计数据空间分析时, 应该对其使用规则格网重新聚合, 并根据实际应用的需求选择多尺度分析方法, 以达到实际应用目的。

关键词: 多边形统计数据; 空间分析; 不确定性; 可塑面积单元问题

DOI: 10.3724/SP.J.1047.2013.00369

1 引言

普查数据通常是以行政区为单元, 通过普查、抽样等方式逐级汇总得到的典型统计型数据^[1]。在地理信息系统中, 该数据一般是作为行政区多边形对象的属性数据进行存储。因此, 本文将该类统计数据称为多边形统计数据, 其在地理学及社会科学研究中较易获得, 可得到广泛使用。该类数据的空间分析较好地揭示了研究对象的空间分布特征^[2-3], 可服务于政府及企业的战略决策工作^[4]。因此, 对于多边形统计数据空间分析的研究方法、评价手段及应用模式的分析, 具有重要的研究意义和科学价值。本文以多边形统计数据空间分析的不确定性, 研究多边形统计数据的可塑面积单元问题效应对空间分析结果的影响模式, 从而对多边形统计数据

空间分析做出评价。

多边形数据空间分析通常以空间统计学为基础, 很大程度上与空间数据的描述与探索有关。由于数据本身不满足经典统计学独立性的假设, 许多情况下, 经典假设检验方法不适用多边形统计数据空间分析^[5]。经过多年的研究, 多边形统计数据的空间分析方法逐渐发展成为描述性、空间统计的两大类核心分析方法^[6]。这两类方法均依赖多边形的距离、方向、形态特征、邻近关系等多边形自身的空间特征。其中, 描述性方法通常为对多边形的属性数据表进行简单的单元统计, 如总值、标准差、均值等, 及对属性数据的可视化直观表达。在空间统计分析中, 统计区(多边形)组成统计区集合 P , P 中每一个统计区 p_i 具有统计指标 z_i ; P 的邻接矩阵 $W=\{w_{ij}\}$ 表达了多边形之间的邻接关系。 $w_{ij}=1$ 表示统

收稿日期: 2012-07-10; 修回日期: 2012-12-24.

基金项目: 国家科技支撑计划项目(2011BAH06B03)。

作者简介: 张小虎(1986-), 男, 江苏宝应人, 博士生, 研究方向为GIS软件技术与统计地理信息系统。

E-mail: zhangxh@lreis.ac.cn

计区 p_i 与统计区 p_j 相邻, $w_{ij}=0$ 表示统计区 p_i 与统计区 p_j 不相邻。多边形统计数据的空间统计分析就是基于多边形邻接关系 W 及相应统计指标的统计分析方法,具体有空间自相关分析与空间集聚分析等。

在地理学及社会科学研究中,不确定性(uncertainty)是一个抽象概念,其含义比误差(error)更为广泛^[7-8],既包含随机误差、系统误差及粗差,也包含数值概念上的误差。多边形统计数据是现实的一个抽象表达,不可避免地存在对所表达的现实特性的不确定性。这种不确定性可以由多边形空间位置的不确定性、拓扑不确定性,及属性的不确定性等引起^[9],也可以由空间分析采用的分析方法导致^[10]。对多边形统计数据而言,针对空间数据质量本身,以多边形顶点位置坐标精度误差去衡量多边形空间误差及相应的属性数据与实测数据不相符的情况。然而,在实际应用中,多边形统计数据作为唯一可靠的研究数据,研究中其不确定性却往往被忽视,主要原因^[7]:(1)缺乏数据。在多数研究中,多边形统计数据是唯一可获得的研究数据,因此,以往的研究只能采用不同等级行政区划数据对该问题进行一般的说明。(2)计算机处理能力的限制。对多边形统计数据空间不确定性研究需要很强的图形及数据处理能力,需要相应的计算机硬件和软件的支持。(3)对于多边形统计数据空间分析的不确定性研究尚未形成成熟的方法。因此,本文提出了一种基于多边形空间特征的多边形统计数据空间分析不确定性研究方案。

2 空间分析不确定性研究方案

多边形空间特征对空间分析的影响主要表现在数据的边界问题及多边形划分问题上,其中多边形划分的影响最为突出^[11]。对于整个研究区,多边形(统计区)主要是按照自然人文要素因子进行划分。而在地理学领域,诸如,气候带、城市带等自然人文现象的边界都是渐变的、逐步过渡的,其边界的划分往往是不确定的。同时,在实际统计数据采集中,多边形对象的选择和划分往往是任意的、可修改的,并且完全取决于人的主观臆断或想象^[7]。不同多边形划分形成对区域内个体单元不同的聚集,其结果反映了具体的多边形区域特征而非个体特征^[13]。在地理学和生态学的研究中,将这种多边形

形的划分问题称为可塑面积单元问题(Modifiable areal unit problem, MAUP)^[12-13]。其具体表现为尺度问题(Scale Problem)与分区问题(Aggregation Problem),不同尺度相同分区(形状)及统一尺度下不同的分区(形状)均会引起该问题。图1给出了一个示例,展现了MAUP对均值的影响。

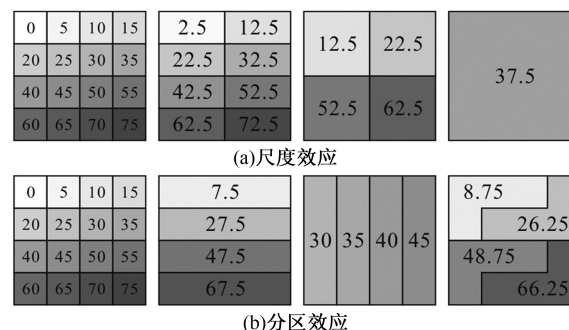


图1 可塑面积单元问题(MAUP)示例图

Fig.1 An example of MAUP: (a) describes scale problem and (b) describes aggregation problem

可塑面积单元问题虽然在地理学研究中被关注了多年^[14-15],但仍然没有一个较为完善的解决方案^[16]。鉴此,OpenShaw提出了最优区划的“解决方案”^[17],然而是否存在最优尺度及最优分区仍然是该研究领域内的一个争论点。同时,Openshaw在前人研究的基础上总结了学界对MAUP的3个不同的基本认识:(1)MAUP是一个无法解决的问题;(2)MAUP是一个可以忽视的问题;(3)MAUP是一个非常有效的分析工具。我们认为,MAUP是空间多边形对象的空间属性赋予地理实体在认识上的客观表现,空间多边形对象的尺度及形态势必引起多边形对象内涵信息的改变。MAUP可以对多边形统计数据的内在信息进行重新认识和展现,因此,MAUP是多边形统计数据空间分析很好的评价手段。李海萍在其研究中指出MAUP对统计数据的空间统计分析存在影响^[11]。为此,本文设计了基于MAUP的多边形统计数据空间分析的不确定性研究方案,从尺度问题和分区问题两个方面,研究多边形数据空间分析的不确定性及其可塑面积单元对人口普查统计数据空间分析的影响模式。

3 多边形空间分析数据与方法

3.1 实验数据

本文采用了2008年全国第二次经济普查时制

作的北京市海淀区建筑物空间单元统计数据作为原始统计区集合。该数据在2010年全国第6次人口普查中被修正完善,共包含107 864个建筑物多边形。由于以建筑物为统计单元的人口数据可近似作为人口统计的总体,因此,此数据构建的不同格网区划人口数据使得探究可塑面积单元对多边形统计数据空间分析不确定性的影响模式成为可能。实验采用北京海淀区2010年11月1日零时的常住人口数作为人口统计指标,在107 864个建筑物中的常住人口为3 216 646人。值得注意的是,官方发布的北京市海淀区常住人口统计数据为3 280 670人^[18],与实验数据有1.9%的差别,这主要是由于建筑物数据库并未完善,一定量的人口与建筑物仍未能关联,但该差异并不影响本文实验。

3.2 实验方法

多边形空间分析的描述性指标和空间统计指标均受可塑面积单元的影响。由于多边形数据MAUP效应的基本机理并未有准确物理模型。本实验采用了模拟研究的方法,即通过模拟可塑面积单元问题的尺度、分区指数和部分空间分析指标的关系,达到研究多边形统计数据空间分析不确定性的目的(图2)。实验的基本步骤为:(1)以建筑物单元的人口普查数据,对数据的谬误及奇异值进行预处理修正;(2)构建形状一致尺度不同的正方

形格网体系及尺度一致形状不同的分区体系;(3)构建不同格网区划,采用了面插值^[19]的方式对修正后的人口统计数据重新聚合,形成格网区划人口统计数据;(4)格网区划人口统计数据,以SuperMap地理信息系统软件为平台,采用可视化分析与数据拟合方法,研究尺度效应及分区效应对多边形统计数据空间分析的影响模式。

3.2.1 实验格网体系的建立

本文构建了两套格网体系分别探索尺度效应和分区效应对多边形统计数据空间分析的影响模式。针对尺度效应,实验采用了统一的规则正方形格网控制分区效应的影响,并采用了5个不同尺度(表1)。为了去除格网数据受边界问题的影响,实验中的格网完全构建在海淀区行政边界内部。

针对分区效应,实验统一采用10 000m²的尺度单元,分别采用规则正方形格网结构,规则矩形格网,规则三角网结构和北京市统计小区多边形^①,研究多边形形态特征与空间分析指标的关系(图3),以此探讨分区效应的影响模式。

3.2.2 多边形统计数据空间分析指标及不确定性评价方法

实验选取了多边形统计数据空间分析的5个指标,包括总值、均值、标准差、全局空间自相关系数(Global Moran's I)及空间聚类分析(Anselin Local Moran's I, LISA)。其中,总值、均值、标准差是描述

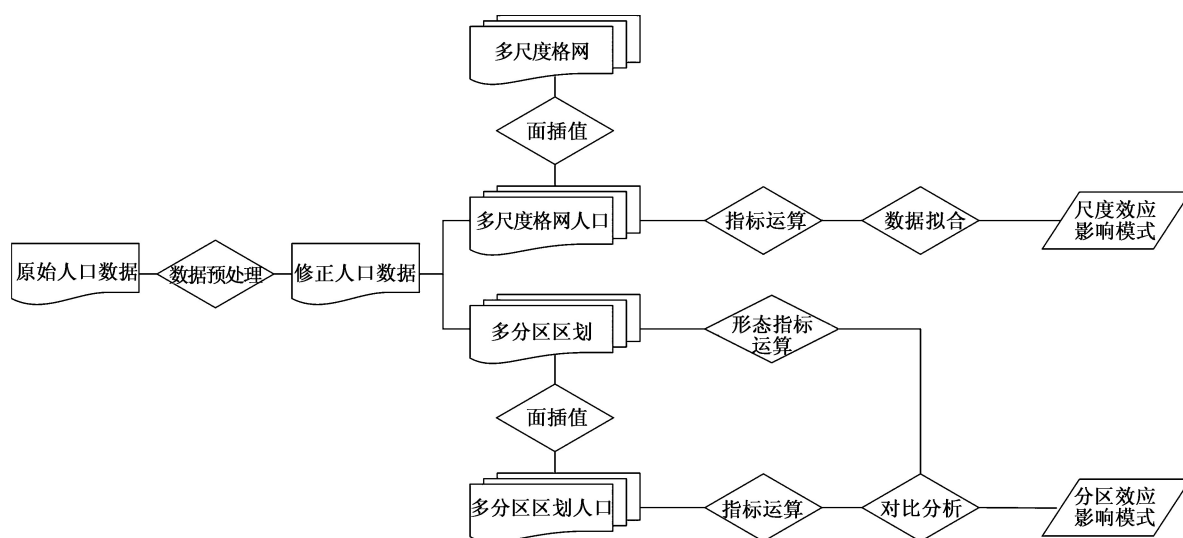


图2 多边形统计数据空间分析的不确定性分析实验流程

Fig.2 Diagram of research on uncertainty of polygon-based statistical data spatial analysis

①北京市海淀区实际统计小区的尺度并非10 000m²,实验中加入实际统计小区以更好说明实际应用状况

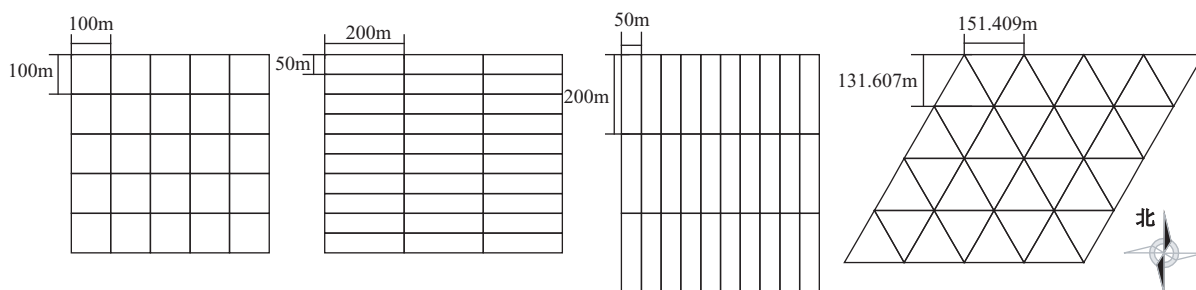
图3 分区效应实验中采用的4种不同形状的格网(10 000m²)Fig.3 Four zoning systems (10 000m²) applied in aggregation effect exploration

表1 尺度效应实验采用的5种不同尺度的规则格网
Tab.1 Five regular grid systems applied in scale effect exploration

	尺度(m)	原始多边形数量	去除边界后多边形数量
1	50×50	174 130	141 600
2	100×100	43 956	35 400
3	250×250	7222	5564
4	500×500	1887	1416
5	1000×1000	505	354

性指标,是对统计数据的概括性度量。全局空间自相关系数和空间聚类分析是空间统计指标。空间自相关系数 Global Moran's I 定义如公式1所示。

$$I = \frac{1}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2} \quad (1)$$

其中, \bar{z} 是 z 的均值。Global Moran's I 反映的是多边形统计数据的整体空间依赖程度,其置信度依靠 P 值和 Z 值检验^[20]。 P 值显著且 Z 值为正值表示空间自相关性强,且 Z 值越大,空间自相关性越显著。LISA 的定义如公式2所示。

$$I_i = \frac{z_i - \bar{z}}{S_i^2} \sum_{j=1, j \neq i}^n w_{ij} (z_j - \bar{z}) \quad (2)$$

其中,

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^n w_{ij}}{n-1} - \bar{z}^2 \quad (3)$$

LISA 的计算结果为多边形的局部集聚模式,分别为高高集聚(HH)、低低集聚(LL)、高低集聚(HL)及低高集聚(LH)。这些模式只是一个相对指标,分析的精确性需要其 Z 值得分来评价^[21]。

实验采用可视化分析和数据拟合两种方法研

究可塑面积单元对指标不确定性的影响模式。这里,可视化分析是指对不同格网区化的人口数据做分级设色分析,直观上阐述了不同格网区划数据表现能力和表现重点。对于尺度效应,实验中尺度以规则格网的边长及面积来表征。采用多项式函数拟合和幂函数拟合的方法展现尺度和空间分析指标的数量关系,其拟合精度用拟合的均方误差(MSE)来衡量。通过评价拟合精度阐述和评价尺度效应对空间分析的不确定性模式。对于分区效应,实验中使用多边形形态指数 P ($P = \text{周长}^2 / \text{面积}$) 及邻接多边形数目 N 表征分区多边形的形态特征。采用对比分析 P 、 N 和空间分析指标的方法阐述分区效应带来的空间分析不确定性模式。

3.3 实验结果及分析

3.3.1 尺度效应导致的不确定性

(1) 尺度效应对数据的直观影响

海淀区人口分布(图4)呈现从市中心(东南)向远郊区(西北)减少的规律,这一总体规律在1km的格网上表现尤为明显,而50m的格网对该信息的表现最弱。50m的格网分辨率高,较好地展现了人口分布的细节,但由于数据特征被碎化,使得全局信息特点被掩盖;1km格网可较好表现数据的总体特征,但由于过量的均值过滤作用,数据局部信息被弱化,从而导致部分原始数据信息丢失。这些直观表现反映了尺度效应对统计数据的影响,不同尺度展现了数据不同的空间特征信息。据此,可以推论不同尺度的多边形数据空间分析的结论也不尽相同。

(2) 尺度效应对空间分析的影响模式

5种不同尺度规则格网化数据的总值、均值、标准差及全局空间自相关系数(Global Moran's I)计算结果如表2所示。数据表明随着格网面积的增大,

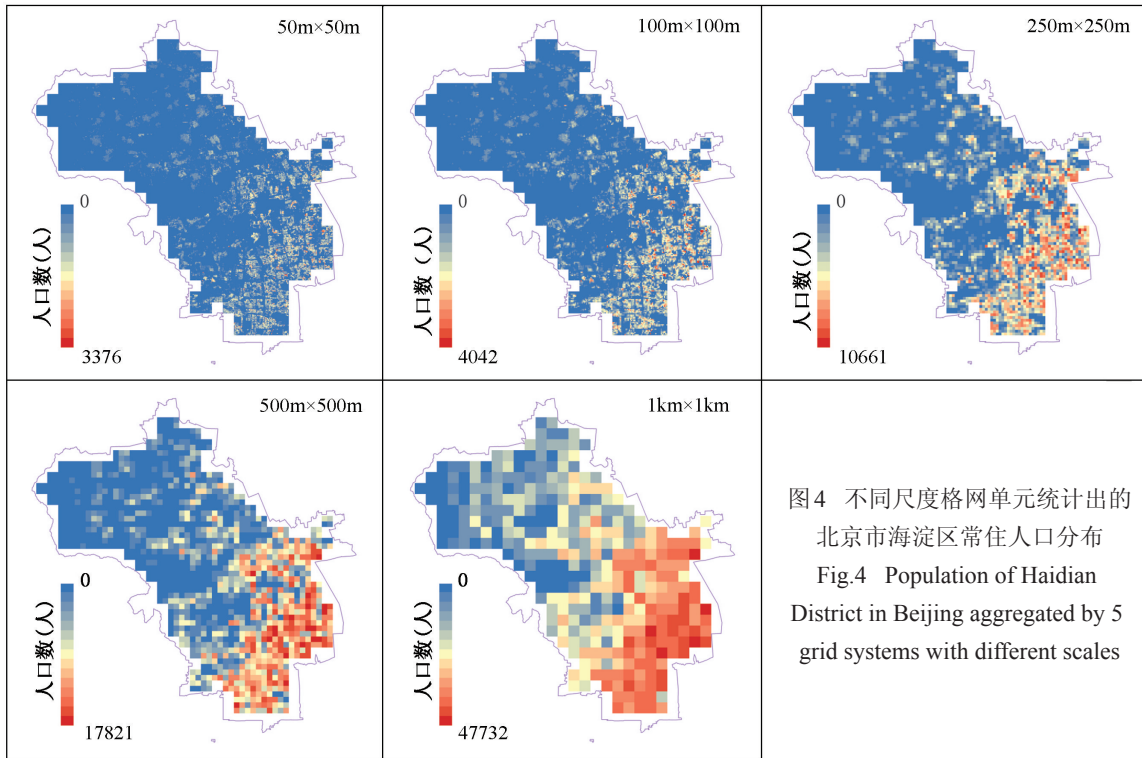


图4 不同尺度格网单元统计出的北京市海淀区常住人口分布
Fig.4 Population of Haidian District in Beijing aggregated by 5 grid systems with different scales

表2 不同尺度人口分布特征空间统计分析指标

Tab.2 Statistical indices calculated on population datasets with different scales

格网(m)	总值	均值	标准差	Moran's I		
				Moran's I	Z值	P值检验
50×50	2546 020.65	17.98	64.57	0.464820	247.027	0.00
100×100	2546 020.68	71.92	193.28	0.532423	141.174	0.00
250×250	2546 020.61	449.51	877.04	0.574074	60.535	0.00
500×500	2546 020.58	1798.04	2836.49	0.632135	33.025	0.00
1000×1000	2546 020.58	7192.15	9530.79	0.707016	18.212	0.00

格网人口数均值变大,标准差变大,总量保持稳定。同时,Moran's I值随着格网尺度的变大而增加,这说明了随着尺度的增大,海淀区人口分布的正空间自相关性逐渐增强,人口的总体集聚性增强。而Z值随着尺度变大而减小,说明这种集聚性虽然显著(0.01 显著水平),但是显著性随着尺度的增大逐渐减弱。

空间分析指标与尺度的数据拟合结果如图4所示。在规则格网下,均值与尺度(面积)呈线性关系,且标准差也和尺度(面积)呈线性关系。从Global Moran's I的定义看(公式1),其主要使用的基本统计指标为统计均值。因此,通过均值与尺度的关系可以推出Global Moran's I与尺度的存在对应的关系。从图5的拟合结果可以看出,Global Moran's I与尺度呈幂函数关系,幂函数指数小于

1;其Z值也与尺度呈幂函数关系,幂函数指数小于1。这表明在规则的正方形格网条件下,统计数据的空间分析指标与尺度呈一定函数关系。同时,均值、标准差和Global Moran's I与面积的拟合精度MSE近似等于1,这表明尺度效应并不对这些指标的分析产生准确性的影响,不同尺度格网划分下的研究结果的不确定性依赖数据在不同尺度下的特征。因此,可以通过尺度将这些研究指标进行统一,避免其不确定性在进一步的分析中造成障碍。

为了探讨尺度效应对局部指标的影响模式,实验研究了尺度对空间聚类分析Anselin Local Moran's I的影响模式(图6)。5种尺度的格网数据均展现了北京市海淀区常住人口在该区东南部有明显的高值集聚现象。Anselin Local Moran's I的Z值均值、标准差与尺度指标的拟合关系表明:随着规则

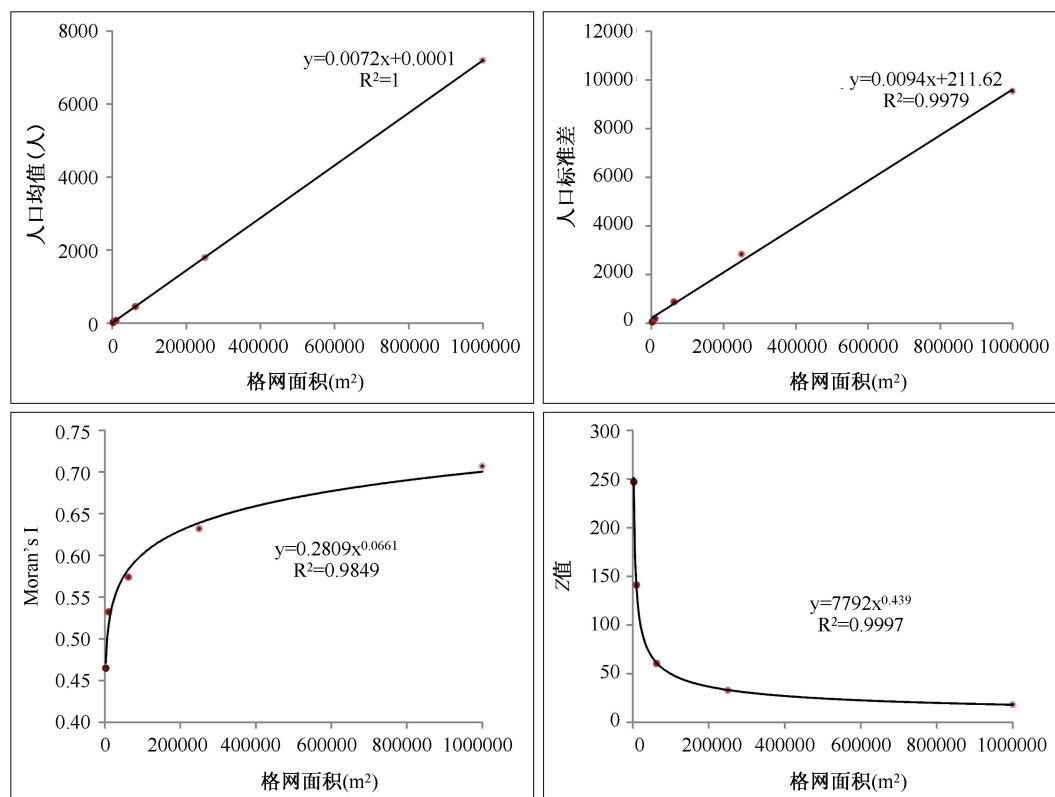
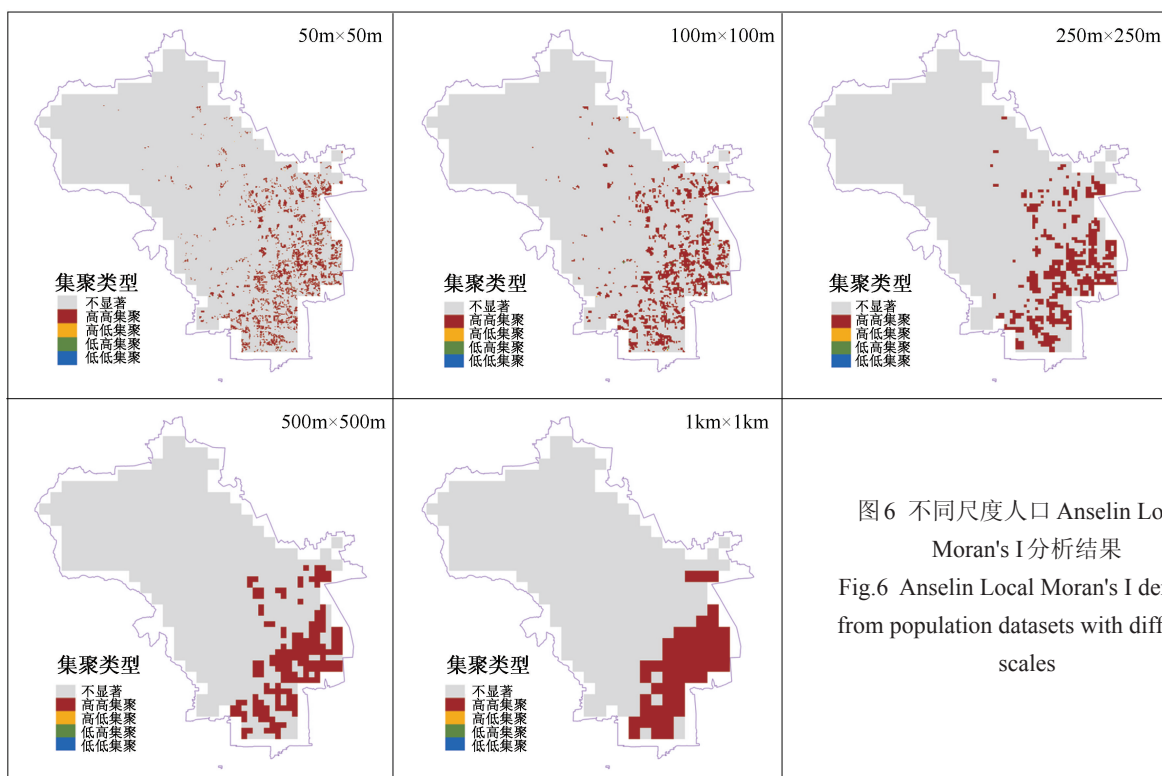


图5 不同尺度人口分布空间特征统计分析指标与尺度关系

Fig.5 Relationship between statistical indices and scale of the grid system

图6 不同尺度人口 Anselin Local
Moran's I 分析结果Fig.6 Anselin Local Moran's I derived
from population datasets with different
scales

格网的尺度增大, 该分析的准确性增强, 分析精度增加。同时, 随着规则格网的尺度增大, 尺度对 Anselin Local Moran's I 的影响能力变小(图7)。

上述实验结果表明, 不能简单使用某一尺度下的格网化数据判断研究对象的空间分布模式, 也不存在描述原始数据空间形态的最优尺度。不同尺度反映的信息量不同, 空间分析结论也不完全一致。在具体的应用中, 才能客观地展现原始数据的空间特征^[22]。

3.3.2 分区效应导致的不确定性

(1) 分区效应对数据的直观影响

实验采用了图3中的4种区划和实际统计小区, 对北京市海淀区建筑物内常住人口进行重新聚合, 结果见图8。规则区划的统计数据可以直观地展现海淀区常住人口从东南向西北递减的分布规律。而实际统计小区的统计数据掩盖了这条规律。这直观表明了不规则区划的分区效应最为强烈, 重新聚合的数据不能很好展示原始数据的空间分布特征。

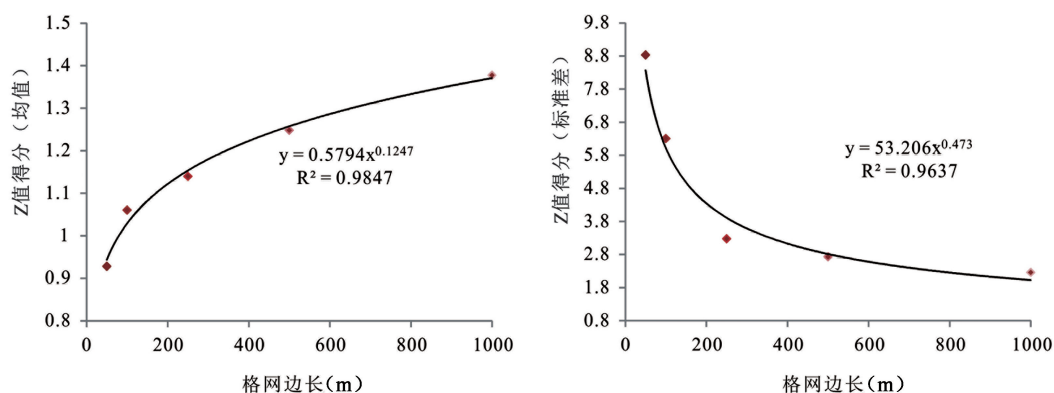


图7 不同尺度人口 Anselin Local Moran's I 分析结果与尺度关系

Fig.7 Relationship between Anselin Local Moran's I and scale of the grid system

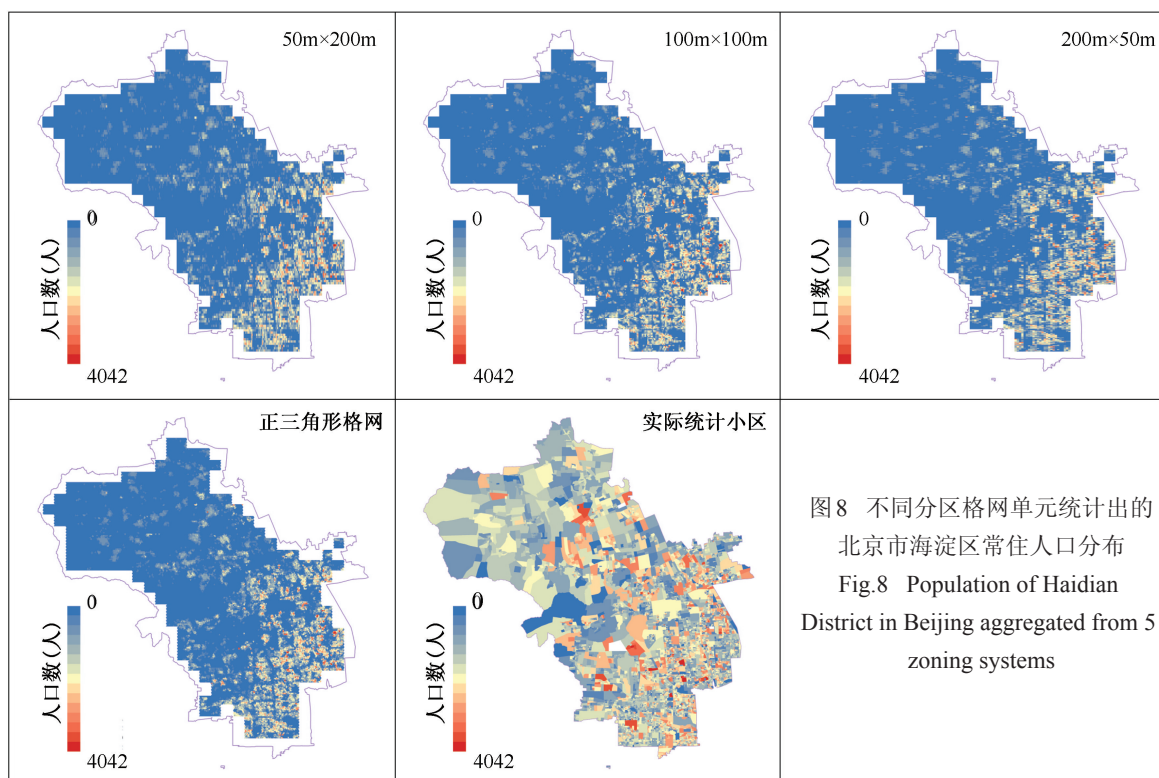


图8 不同分区格网单元统计出的
北京市海淀区常住人口分布

Fig.8 Population of Haidian
District in Beijing aggregated from 5
zoning systems

(2) 分区效应对空间分析的影响模式

5种格网数据的总值、均值、标准差及全局空间自相关系数(Global Moran's I)如表3所示。其中,由于正三角形与矩形的边界不一致,正三角形区划的总面积不同于矩形数据,因此,其总值没有参考意义。指数P、指数N与分析指标对比关系表明规则格网区划的分区效应体现在原始数据分布形态

与格网形状的关系上。随着P指数增大,统计数据的标准差减小,即规则格网区划下,格网越偏离圆形(简单欧几里得形状),数据之间的差异越明显。同时,具有相同P指数而方向不同的标准格网,对数据的反映也不尽相同。其原因主要是由于原始建筑物内人口数据的空间分布并非均质,格网的不同方向强化或弱化了该特征。

表3 不同分区人口分布特征空间统计分析指标

Tab.3 Statistical indices calculated on population datasets with different zoning systems

	P(周长 ² /面积)	邻接多边形数	均值	标准差	Moran's I		
					Moran's I	Z值	P值检验
50m×200m	25	8	71.92	186.77	0.533596	141.107	0.00
100m×100m	16	8	71.92	193.28	0.532423	141.174	0.00
200m×50m	25	8	71.92	190.41	0.505244	133.662	0.00
正三角形	21	12	71.91	192.00	0.567787	128.757	0.00
统计小区	20	6	—	—	0.308692	49.373	0.00

同时,规则格网形状的差别导致其邻接多边形数目不同,从而影响计算分析时每一个多边形实际的计算尺度:对于矩形格网,计算尺度为9个网格共90 000 m²,而三角形格网的计算尺度为130 000 m²。根据尺度效应的分析结论,正三角形区划的Moran's I的计算值偏大,显著性较低。因此,分区效应带来的不一致性对于规则格网区划体现在格网的形状和方向上,并受到分析方法的实际计算尺度影响。对于实际的统计小区,其几何特征更为复杂,并且尺度不统一,分析结果的不确定性受尺度与分区效用的共同影响,从而表现得更强。这表明当区划数据成为研究对象时,区划的空间属性赋予了地理实体数据空间特征。越复杂的区划,其空间特征的权重越强。数据受到分区效应的影响也就越强,空间分析的结论越不可靠。

为了探讨分区效应对局部指标的影响模式,实验分析了分区效应对Anselin Local Moran's I的影响模式(图9)。对于规则格网区划,整个研究区的集聚特征基本一致,海淀区东南部高值集聚;由于尺度较小,细节信息反映较为明显。LISA显著性如表4所示。结果表明,同样尺度、不同的邻接多边形数导致不同的分析结果。对于规则格网,局域分析结果的不确定性受多边形邻接多边形数目影响;相同领域数目的格网其结果大致相当(矩形格网)。同时,实际统计小区的统计数据的结果受分区形式影响较大,没有很好地展示原始数据的

集聚特征,需要采用合适的方法将其离散成格网加以使用^[23]。

表4 不同分区人口 Anselin Local Moran's I 分析结果与分区关系

Tab.4 Relationship between Anselin Local Moran's I and zoning systems

	邻接多边形数	高值集聚多边形数*	Z值得分均值	Z值得分标准差
50m×200m	8	2944	1.0618	5.697
100m×100m	8	2938	1.0605	6.304
200m×50m	8	2914	1.0049	6.003
正三角形*	12	2675	0.9796	5.957

*总格网数为35 400

4 结论与讨论

本文采用北京市海淀区建筑物内常住人口统计数据,以可塑面积单元问题的相关方法,分析了多边形统计数据空间分析的不确定性。实验设计了不同尺度的规则格网和同尺度下不同分区规则格网区划,分别就尺度效应和分区效应对多边形统计数据空间分析的影响模式做了探究分析。研究结果表明MAUP对多边形统计数据的使用存在巨大影响,并存在一定的模式,主要结论如下:

(1)规则格网对原始数据空间形态的反映更为真实。本质上,当以区划数据为研究对象时,区划

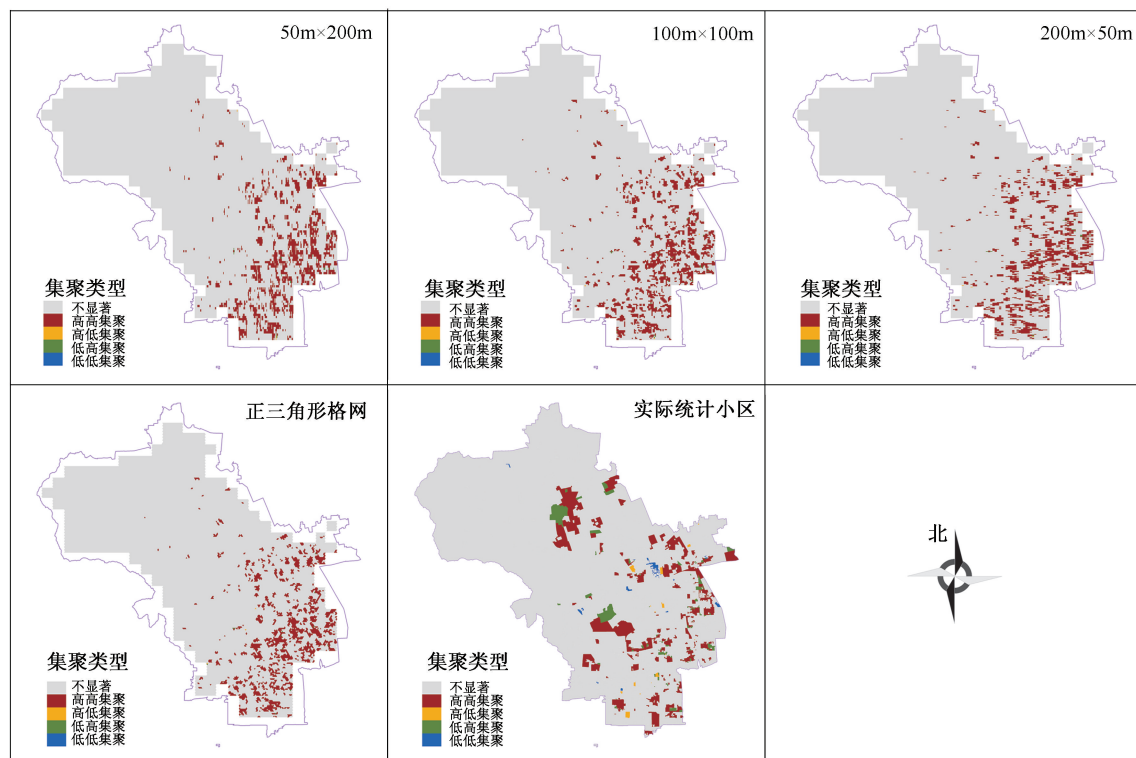


图9 不同分区人口 Anselin Local Moran's I 分析结果

Fig.9 Anselin Local Moran's I derived from population datasets with different zoning systems

的空间属性赋予了地理实体数据空间特征,同时在区划数据上反映出来,复杂的统计小区形态给予了统计数据更多的小区空间特征信息,弱化甚至掩盖了实际数据的分布特征。因此,建议使用规则格网多边形削弱这种影响,而规则的栅格便是一个比较合适的多边形格网形式。

(2)规则格网数据的空间统计分析结果受尺度的影响,同时分析结果与尺度有较好的拟合关系,对于不同尺度下的规则格网数据分析结果可以使用多尺度信息加以综合,避免对进一步使用分析结果造成障碍。

(3)对于规则格网,不存在最优尺度,不同的尺度反映的是原始数据不同的信息。尺度作为空间对象内含的要素,其考量值发生变化时,其全局空间形态及相应的空间分析结果均产生变化,并给出相应尺度上的正确信息。因此,对于统计数据而言,不同尺度的综合分析才是最佳分析方法。

(4)对于规则格网,局域分析的显著性受到其尺度的影响,然而当尺度增大到一定程度,尺度的影响能力变小,数据分析结果准确性趋于稳定。因此,结果尺度是否适用于具体应用成为实际应用中应该注意的问题,而非一味追求尺度的最优。

(5)规则格网的分区效应受尺度效应影响,并共同影响空间分析结果。格网的不同形状对应了格网不同的空间形态,对数据进行了不同方向和范围的概括。同时,不同形状导致了格网不同的邻域范围,从而使得空间分析的实际计算范围不一致,引起分析结果差异。相同尺度下具有相同领域数目的格网数据的分析结果相当。

(6)实际统计小区的数据包涵了丰富的统计小区空间形态信息,这些信息弱化了原始数据的实际空间分布特征,并在数据及分析上表现出来。因此,不规则格网(实际分化小区)实验结果的空间分析结果不适合作为参考的标准结论加以使用。

上述可见,在实际使用多边形统计数据(如年鉴数据)时,需要将其离散成标准的格网形态,并采用多个尺度对数据进行预处理,才能获得有参考价值的空间分析结果。

参考文献:

- [1] 廖顺宝,李泽辉.基于人口分布与土地利用关系的人口数据空间化研究——以西藏自治区为例[J].自然资源学报, 2003,18(6):659-665.
- [2] 陈浩,邓祥征.中国区域经济发展的地区差异GIS分析[J].地球信息科学学报,2011,13(5):586-593.

- [3] 董冠鹏,郭腾云,马静.京津冀都市区经济增长空间分异的GIS分析[J].地球信息科学学报,2010,12(6):797-805.
- [4] 安凯,陈炎平,张锦水,等.甘肃省统计地理信息系统建设研究[C].中国地理信息系统协会第三次代表大会暨第七周年会,2003.
- [5] 杜培军,张海荣,冷海龙.地理空间分析——原理、技术与软件工具[M].北京:电子工业出版社,2009.
- [6] Cressie N. Statistics for spatial data[J]. Terra Nova, 1992,4(5):613-617
- [7] Openshaw S. The modifiable areal unit problem[M]. Concepts and Techniques in Modern Geography. Norwich: Geo Book,1984.
- [8] Goodchild M F. Issues of quality and uncertainty[M]. New York: Elsevier,1991.
- [9] 郭伦,于海龙,高振纪,等.GIS不确定性框架体系与数据不确定性研究方法[J].地理学与国土研究,2002,18(4):1-5.
- [10] 史文中.空间数据与空间分析不确定性原理[M].北京:科学出版社,2005.
- [11] 李海萍.空间统计分析中的MAUP及其影响[J].统计与决策,2009(22):15-17.
- [12] Gehlke C, Biehl K. Certain effects of grouping upon the size of the correlation coefficient in census tract material [J]. J Am Stat Assoc, 1934,29(185):169-170.
- [13] 郭建国,JELINSKI D. 生态学中的格局与尺度-可塑性面积单元问题[M].北京:科学出版社,1995.
- [14] Openshaw S, Taylor P. A million or so correlation coefficients statistical methods in the spatial sciences[M]. London: Pion,1979,127-144.
- [15] Fotheringham A S, Densham P J, Curtis A. The zone definition problem in location - Allocation modeling[J]. Geogr Anal,1995,27(1):60-77.
- [16] Fotheringham A S, Brunsdon C, Charlton M. Quantitative geography: Perspectives on spatial data analysis[M]. London: Sage Publications Ltd,2000,237-240.
- [17] Openshaw S, Rao L. Algorithms for reengineering 1991 census geography[J]. Environ Plann A, 1995,27(3): 425-446.
- [18] 海淀区第六次全国人口普查领导小组办公室. 海淀区2010年第六次全国人口普查主要数据公报(1)[OL].海淀区统计局. 2011.
- [19] Goodchild M, Anselin L, Deichmann U. A framework for the areal interpolation of socioeconomic data[J]. Environ Plann A,1993,25(3):383-397.
- [20] Moran P A P. Notes on continuous stochastic phenomena [J]. Biometrika, 1950,37(1/2):17-23.
- [21] Anselin L. Local indicators of spatial association—LISA [J]. Geogr Anal, 1995,27(2): 93-115.
- [22] 杜国明,张树文,张有全.城市人口分布的空间自相关分析——以沈阳市为例[J].地理研究,2007,26(2):383-390.
- [23] Zhang X, Zhong E, Zheng H, *et al.* Reconstructing continuous population density surface from polygon-based data [C]. IEEE, 2010.

The Uncertainty of Polygon-based Statistical Data Spatial Analysis: Case of Census Data of Haidian District, Beijing

ZHANG Xiaohu^{1,2*}, ZHONG Ershun¹, WANG Shaohua^{1,2}, ZHANG Xun^{1,2} and ZHANG Ji³

(1. *Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China*; 2. *University of Chinese Academy of Sciences, Beijing 100049, China*; 3. *Planning and Design Institute of Forest Products Industry, State Forestry Administration, Beijing 100714, China*)

Abstract: In statistic geographic information system, census data, stored as polygon attribute, is a kind of polygon-based statistical data. Moreover, in the studies of geography and social science, polygon-based statistical data is a main data source for uncovering spatial patterns of social phenomena by spatial analysis. However, due to the limitation of data and restriction of computer processing power, uncertainty of polygon-based statistical data spatial analysis is always ignored, and there is no well methodology for analyzing such uncertainty. To address this question, we developed a method concerning modifiable areal unit problem (MAUP) to evaluate uncertainty of polygon-based statistical data spatial analysis. The population data collected from each building in Beijing makes the method applicable. For MAUP, we considered it as scale and aggregation separately. For polygon-based statistical data, we applied census data of Haidian District (Beijing) with polygons of buildings as its

georeference. With this method, we introduced scale and shape indices and applied visual analysis and data fitting to detect the uncertainty of five analysis methods: Sum, Mean, Standard deviation, Global Moran's I and Anselin Local Moran's I (LISA). In addition, the relationships between scale, shape indices and the five analysis methods are also revealed in order to demonstrate the way that MAUP affects polygon-based statistical data spatial analysis. The result of the research shows as follows: (1) the results derived from census data spatial analysis with normal census tracts as zone system are arbitrary and have great uncertainty. (2) The results derived from census data spatial analysis with regular nets as zone system well describe the spatial patterns of original data, but still depend on the scale and zoning of the net system. (3) The results derived from census data spatial analysis with regular grid as zone system, are functionally related to the scale of the grid system, and the uncertainty of the results represents multi-scale spatial patterns of original data. And (4) aggregation together with scale affects census data spatial analysis. With regard to regular net system with fixed scale, the number of the neighbors of each polygon affects the results of the analysis. According to the above, it is better to re-aggregate the census data by regular grid system with proper scale and apply multi-scale methods in polygon-based statistical data analysis.

Key words: polygon-based statistical data; spatial analysis; uncertainty; modifiable areal unit problem

***Corresponding author:** ZHANG Xiaohu, E-mail: zhangxh@lreis.ac.cn