文章编号: 1671-8860(2006)03-0240-04 文献标志码: A

顾及空间自相关的统计数据分级质量评价

郭庆胜[1,2] 李留所[2] 贾玉明[2] 孙 艳[1,2]

(1 武汉大学地理信息系统教育部重点实验室, 武汉市珞喻路129号, 430079)
(2 武汉大学资源与环境科学学院, 武汉市珞喻路129号, 430079)

摘 要: 详细分析了统计地图数据分级质量的评价指标, 研究了数据分级中应当考虑的数据空间分布规律, 并用实例证明了分级数与空间自相关系数之间的变化规律。

关键词: 统计地图; 数据分级; 空间自相关; 质量评价
中图法分类号: P283; P208

统计数据分级是专题地图综合中的一个重要问题, 其目的是在尽可能少地丢失原始信息的基础上, 将大量的观察数据进行归纳合并, 并且保证数据分级后在统计分布和空间分布上能尽可能反映出现象的本质[1]。国内外已提出了很多种分级方法和分级质量评价指标。在质量评价方面, 目前主要以视觉变量和空间认知为基础, 在保证分级视觉效果的前提下, 根据统计学原理对分级的数据精度进行评价[1-4], 很少考虑数据的空间分布特征。文献[5]把遗传算法用于统计地图的数据分级, 分析了多个质量评价指标两两结合用于评价分级质量的方法, 以便找到多标准下的最优解, 他们也认为在分级中应当考虑数据的地理特征。当然统计数据千差万别, 地理特征很多, 但是在分级过程中, 数据的空间分布规律是必须考虑的, 即不仅要考虑数据的统计精度、图面视觉效果、空间认知, 还要考虑分级区域破碎程度和空间自相关程度等。

1 分级质量评价指标

统计制图中数据分级的主要原则包括: 统计学方面的要求, 即保持数据的主要统计特征; 制图学方面的要求, 即尽量保持原始数据的空间分布特征; 增强地图信息的传输效率。从人的视觉效果和空间认知能力来看, 一般认为4~7级是合适的分级数。这就需要通过相应的评价指标来确定。陆效中1989年提出了6个评价数据分级精度的指标, 指标值越大, 精度越高。这些指标是总偏差误差 EC_1、加权总偏差误差 EC_2、平均偏差误差 EC_3、加权平均偏差误差 EC_4、综合误差 SumEC 和分级匹配精误差 MEC[1]。

原始数据本身就有潜在的空间分布规律, 这种空间分布特征在地理研究中是需要考虑的重要因素。对于以区域为统计单位的数据而言, 相邻区域之间的数据差别和数据的空间自相关程度都是用来描述空间分布特征的非常重要的指标。数据分级后, 区域的属性值就被人为抽象了, 相邻区域之间的差别就会发生变化。同样道理, 空间自相关程度也会变化。也有学者认为, 数据分级后, 各级所占的区域面积应当符合一定的统计规律, 只有当统计数据与统计单元的面积密切相关时, 才需要考虑。例如, 符合正态分布规律; 各级所占的区域面积应当尽量相等, 以便读图者能快速地从地图上获取数据在空间分布上的差异, 也可以从视觉上达到一种平衡。但是, 正态分布规律的保持需要足够数量的样本数, 统计数据所关联的区域面积的统计规律有时也并不符合正态分布规律。笔者认为, 在选择分级方法时需要考虑这个问题。

1.1 区域边界误差 BE(boundary error)

该参数说明了数据分级后, 新的不同级别之间两边的数据差异变化情况。理论上, 空间上相邻的数据如果差别不大, 就应当归为一级, 数据分

收稿日期: 2005-12-31。
项目来源: 国家教育部高校骨干教师专项基金资助项目; 国家自然科学基金资助项目(40571133)。

级后, 区域边界两边的统计值就发生了变化, 当然, 这种变化越小越好, 其表达式如下[5]:

$$BE = 1 - \frac{\sum_{i=1}^{\|H\|}\sum_{(r,l)\in H} | x_{ir} - x_{il} |}{\sum_{i=1}^{\|H\|}\sum_{(r,l)\in G} | x_{ir} - x_{il} |}$$

其中, G 为空间上相邻的区域单元对, 如图 1 所示, 区域单元为 A、B、C、D 和 E, 其中 C、D 为一级, A、B、E 为一级, 则 G 为(A, B)、(A, C)、(B, C)、(C, D)、(C, E)、(D, E); H 为属于不同级别的相邻单元对, 对于图 1 而言, H 为(A, C)、(B, C)、(C, E)、(D, E); $\|H\|$ 为 H 的个数, 图 1 中的 $\|H\|$ 为 4; l、r 分别表示相邻单元对的左单元、右单元; x_{il} 表示不同级别的相邻单元左边单元值, x_{ir} 表示不同级别的相邻单元右边单元值。BE 的取值范围为$[0, 1]$, 在分级数确定的情况下, BE 越小, 表明在地理空间上相邻的不同级别的区域单元差别越大, 分级效果越好。该分级方案很好地反映了事物在地理空间分布上的实际差异。

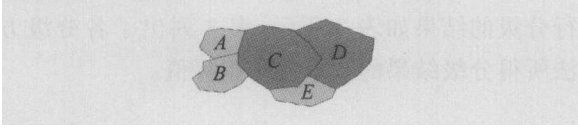

图 1　拓扑关系示意图

Fig. 1　Sketch Map of Topological Relation

1.2　地理面积均等程度 GAE(geographical area equalization)

不同级别之间的地理面积均等程度是从视觉重量感的观点出发, 用于检测各级区域的不一致性。该指标表达式如下(参考文献[5]的表达式有误):

$$GAE = \max\left| \frac{i}{k} - \frac{1}{A}\sum_{j=1}^{i} A_j \right|, i = 1, \cdots, k$$

其中 k 为分级数, A 为所有级别的区域面积之和, $A_1, \cdots, A_k$ 按升序排列, A_j 为第 j 级所包含的区域面积总和。GAE 的取值范围为$[0, 1]$, GAE 越小, 各级区域面积越均衡一致, 分级的视觉效果越好。

1.3　空间自相关系数 MIC(moran's I statistic coefficient)

空间自相关系数是检测邻近单元相似性的重要指标, 可以用于检验空间变量的取值是否与相邻空间上该变量取值大小有关。MIC 取值范围在 -1~1 之间, 当 MIC=0 时, 代表空间不相关, MIC 为正数时, 表明空间变量在一点上的取值与相邻点的取值变化趋势相同, 被称为空间正相关, 相反则被称为空间负相关。空间自相关分析首先要对所检验的空间单元进行配对和采样, 本文对直接相邻的群进行配对, 全部采样[5,6]。

$$MIC = \frac{M\sum_{(i,j)\in C}(x_i - \bar{x})(x_j - \bar{x})}{\|P\|\sum_{i=1}^{M}(x_i - \bar{x})^2}$$

其中, 群为属于同一级的邻近的单元集(在地图上表现为同一级相邻接的单元所构成的多边形)。图 1 中, A 和 B 为一个群, C 和 D 为一个群, E 为一个群, M 为群的个数, P 为邻近群对的集合, $\|P\|$ 为 P 的个数, $\bar{x}$ 为所有群的平均值, x_i、x_j 为两相邻的群 i、群 j 对应的值或所在分级的均值。Armstrong 等(2003 年)认为, 为了使各评价指数的取值范围保持一致, 以方便比较与计算, 规定 $MIC_1 = (1 - MIC)/2$, 这里 MIC_1 的取值范围为$[0, 1]$, MIC_1 越小表明级间空间自相关越大。但是, 原始数据本身就存在着空间自相关性, 若一味追求该值的最小化, 就会与原始数据的空间自相关情况相违背, 因此, 应尽量保持分级后数据的空间自相关程度与原始数据空间自相关程度一致才是合理的。该指标对于划分为一个等级的相邻区域集合的局域空间自相关的评价是比较合理的。

2　MIC_1 与分级数之间的变化规律

以某地区的 31 个县人口数为例, 见表 1[7], 为了简化原始空间数据, 这里以图 2 的子区域为统计单位, 图中的号码代表子区域的统计序号。

表 1　统计数据表

Tab. 1　Data Table

序号	数据	序号	数据	序号	数据	序号	数据
1	202 083	9	346 484	17	479 824	25	588 006
2	232 678	10	354 847	18	500 899	26	609 555
3	264 514	11	357 756	19	503 124	27	610 998
4	280 498	12	369 194	20	512 166	28	820 142
5	301 779	13	373 105	21	517 093	29	951 806
6	312 889	14	374 483	22	534 594	30	958 281
7	320 982	15	390 706	23	541 679	31	1 016 065
8	343 486	16	473 305	24	561 782		

采用系统聚类法将表 1 中的数据逐步从 31 级归类为 1 级, 分级数从 1 到 31 的 MIC_1 的值依次对应为: 0、0.927 47、0.596 46、0.596 46、0.466 7、0.434 1、0.373、0.324 5、0.324 5、0.324 5、0.335、0.328 1、0.328 1、0.325 1、0.325 1、0.317 1、0.317 1、0.317 1、0.317 1、0.317 1、0.365 8、0.317 1、0.318 9、0.308 4、0.399 2、0.300 4、0.300 4、0.300 4、0.275

0.283 6、0.283 6, 分级数所对应 MIC_1 的变化趋势如图 3 所示。

3	2	4	27	31	29
1	6	28	26	17	30
11	12	25	16	19	18
13	23	20	21	8	7
5	15	14	24	22	9
10					

图 2 统计单元

Fig. 2 Statistic Units

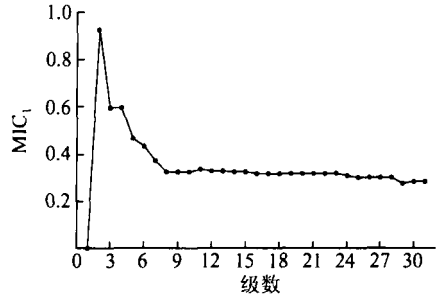

图 3 分级数与 MIC_1 的关系

Fig. 3 Relationship Between Number of Classes and MIC_1

从图 3 可以看出, 级数为 1, 即整个区域分为 1 级时, MIC_1 取最小值 0, 对应空间自相关系数 MIC 取最大值 1, 若只单独考虑该指标的最小化, 那么可以认为, 分级为 1 级的效果最好, 但显然无意义。从分级数为 2 到分级数为 8, MIC_1 迅速变小, 从分级数为 8 到分级数为 31 变化幅度很小, 实际上, 分级数为 31 时的 MIC_1 就说明了原始数据的空间自相关系数。同时, 分级数为 2、3、4 时, MIC 为负值, 是空间负相关, 应该排除; 分级数为 8 级以后, MIC_1 变化很小。因此, 从空间自相关来看, 5~ 8 级是合适的。

3 综合性评价

统计制图中数据分级的方法很多, 这里选择 8 种方法: 最优分割法、逐步聚类分级法、模糊聚类分级法、逐步模糊模式识别分级法、任意数列分级、算术级数分级、几何级数分级、任意级数分级, 以上各方法均分为 7 级[1, 2, 7], 对表 1 中的数据进行分级的结果如表 2 所示, 表 3 列出了各分级方法所得分级结果的 9 个评价指标值。

表 2 不同方法的分级结果

Tab. 2 Results of Different Classification Methods

方法	等级	分级界线	数据个数	方法	等级	分级界线	数据个数
最优分割分级	1	200 000~ 250 000	2	逐步聚类分级法	1	200 000~ 250 000	2
	2	250 000~ 290 000	2		2	250 000~ 330 000	5
	3	290 000~ 330 000	3		3	330 000~ 430 000	8
	4	3 330 000~ 430 000	8		4	430 000~ 530 000	6
	5	430 000~ 540 000	7		5	530 000~ 720 000	6
	6	540 000~ 700 000	5		6	720 000~ 890 000	1
	7	700 000~ 1 020 000	4		7	890 000~ 1 020 000	3
模糊聚类法	1	200 000~ 250 000	2	逐步模糊识别法	1	200 000~ 270 000	3
	2	250 000~ 330 000	5		2	270 000~ 330 000	4
	3	330 000~ 430 000	8		3	330 000~ 380 000	7
	4	430 000~ 490 000	2		4	380 000~ 430 000	1
	5	490 000~ 720 000	10		5	490 000~ 520 000	6
	6	720 000~ 890 000	1		6	520 000~ 720 000	6
	7	890 000~ 1 020 000	3		7	720 000~ 1 020 000	4
任意数列分级	1	200 000~ 240 000	2	算术级数分级法	1	200 000~ 250 000	2
	2	240 000~ 290 000	2		2	250 000~ 330 000	5
	3	290 000~ 340 000	3		3	330 000~ 420 000	8
	4	340 000~ 420 000	8		4	420 000~ 540 000	7
	5	420 000~ 540 000	7		5	540 000~ 670 000	5
	6	540 000~ 720 000	5		6	670 000~ 830 000	1
	7	720 000~ 1 020 000	4		7	830 000~ 1 020 000	3
几何级数分级	1	200 000~ 270 000	3	任意级数分级	1	200 000~ 250 000	2
	2	270 000~ 340 000	4		2	250 000~ 330 000	5
	3	340 000~ 430 000	8		3	330 000~ 430 000	8
	4	430 000~ 540 000	7		4	430 000~ 550 000	8
	5	540 000~ 670 000	5		5	550 000~ 690 000	4
	6	670 000~ 830 000	1		6	690 000~ 850 000	1
	7	830 000~ 1 020 000	3		7	850 000~ 1 020 000	3

表 3　分级评价指标

Tab. 3　Classification Evaluation Index

分级方法	GAE	BE	MIC_1	EC_1	EC_2	EC_3	EC_4	SumEC	MEC
取值范围	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]
逐步聚类	0.221 2	0.106 0	0.373 0	0.113 3	0.126 5	0.104 1	0.119 4	0.115 8	0.037 3
模糊聚类	0.221 2	0.165 9	0.347 8	0.133 4	0.144 2	0.100 8	0.116 7	0.123 7	0.038 8
任意数列	0.345 6	0.202 8	0.367 8	0.131 6	0.122 9	0.129 4	0.129 4	0.128 3	0.042 7
算术级数	0.221 2	0.156 7	0.367 8	0.112 5	0.125 9	0.105 0	0.120 0	0.115 9	0.046 4
几何级数	0.202 8	0.156 7	0.371 4	0.110 3	0.123 9	0.104 7	0.119 1	0.114 5	0.042 2
任意级数	0.221 2	0.170 5	0.360 8	0.112 9	0.126 7	0.105 0	0.119 9	0.116 1	0.042 1
最优分割分级	0.345 6	0.202 8	0.399 2	0.131 6	0.122 9	0.129 4	0.129 4	0.128 3	0.040 7
逐步模式识别分级	0.230 4	0.087 6	0.411 5	0.135 9	0.132 6	0.129 7	0.130 8	0.132 3	0.039 1

前面已分析了各个评价指标的作用,除 MIC_1 外,其他的 8 个指标都是值越小越好,同时 EC_1、EC_2、EC_3、EC_4 这 4 个指标与 SumEC 有关联,SumEC 是它们的平均值。当 MIC 只是局域空间自相关系数时,也可以认为 MIC_1 的值越小越好。从理论上讲,能保证这些评价指标同时最小的分级无疑是最好的分级方法,然而,由于这些指标之间的相互制约性(例如,追求各级多边形的面积均衡,就很难保证各级内多边形群的邻近性),这样的分级实际上是很难存在的。从表 3 来看,根据指标 GAE 进行权衡,几何级数分级的效果最好;根据 BE 指标权衡,逐步模式识别分级法的效果最好。然而,为了保证分级的整体效果最好,就需要兼顾多个指标,综合性评价公式如下:

$$A = a_1\mathrm{GAE} + a_2\mathrm{BE} + a_3\mathrm{MIC}_1 + a_4\mathrm{EC}_1 + a_5\mathrm{EC}_2 + a_6\mathrm{EC}_3 + a_7\mathrm{EC}_4 + a_8\mathrm{SumEC} + a_9\mathrm{MEC}$$

式中,a_1、a_2、a_3、a_4、a_5、a_6、a_7、a_8、a_9 为各指标对应的权重,且 $\sum_{i=1}^{9} a_i = 1$,a_i 取值范围为$[0, 1]$,其取值可根据制图目的、要求并结合经验进行调整。例如,若强调空间分布规律,则可以取值 $a_2 = a_3 = 0.3$,$a_1 = a_4 = a_5 = a_6 = a_7 = 0$,$a_8 = a_9 = 0.2$。其实,用户可以通过人机交互的方式得到满意的结果。

4　结　语

实际操作中,制图者可以在适当兼顾其他指标值相对小的情况下,选择自己特别关注的指标值,获得满意的分级结果,这种评价方法就为制图者提供了较多选择,以便找到更适合的分级。也可以把 MIC_1 单独列出,在统计误差评价方面甚至可以只考虑 SumEC 和 MEC,那么评价指标公式就变为:

$$\begin{cases} |\mathrm{MIC}_k - \mathrm{MIC}_n| \leqslant \varepsilon \\ \min: a_1\mathrm{GAE} + a_2\mathrm{BE} + a_8\mathrm{SumEC} + a_9\mathrm{MEC} \end{cases}$$

式中,MIC_k 是分为 k 级时改正后的空间自相关系数;MIC_n 是 n 个原始数据的改正后空间自相关系数,ε 为给定的阈值,$a_1 + a_2 + a_8 + a_9 = 1.0$。

参　考　文　献

[1] 陆效中. 统计地图的分级表示法[M]. 北京: 解放军出版社, 1989

[2] 黄仁涛, 庞小平, 马晨燕. 专题地图编制[M]. 武汉: 武汉大学出版社, 2003

[3] 祝国瑞, 郭礼珍, 尹贡白, 等. 地图设计与编绘[M]. 武汉: 武汉大学出版社, 2001

[4] 李铭. 专题地图统计数据分级的模式识别方法的研究[M]. 常德师范学院学报(自然科学版), 2000, 12(1): 78-81

[5] Armstrong M P, Xiao Ningchuan, Bennett D A. Using Genetic Algorithms to Create Multicriteria Class Intervals for Choropleth[J]. Annals of the Association of American Geographers, 2003, 93(3): 595-623

[6] 曾辉, 江子瀛, 孔宁宁, 等. 快速城市化景观格局的空间自相关特征分析——以深圳市龙华地区为例[J]. 北京: 北京大学学报(自然科学版), 2000, 36(6): 824-831

[7] 何宗宜. 用信息论方法确定地图分级[J]. 四川测绘, 1995, 18(1): 18-22

第一作者简介: 郭庆胜, 博士, 教授, 博士生导师。主要从事地图制图综合、地理信息智能化处理与可视化研究。

E-mail: guoqingsheng@yahoo.com

(下转第 251 页)

the following edge(connected line) turned relative to the current are founded. By deriving the direction relationship between the neighbouring edges(connected lines) based on these principles, the polygon's orientation, polygonal convexity-concavity can be identified, so does to point-in-polygon query. New algorithms are presented, the analysis shows that improved performance is achieved, and there is the unification of the geometry idea in solving those four issues.

Key words: polygon; orientation identification; convexity-concavity identification; point-in-polygon query; Q_i operator

About the first author: DING Jian, Ph. D, lecture, majors in GIS and its application in field water supply.
E-mail: ydjian@yahoo.com.cn; yudingjian@sina.com.cn

(上接第 243 页)

Quality Evaluation of Statistical Data Classification Considering Spatial Autocorrelation

GUO Qingsheng[1,2]　*LI Liusuo*[2]　*JIA Yuming*[2]　*SUN Yan*[1,2]
(1 Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)
(2 School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: Statistical map is the kind of important forms of thematic map, and data classification is the kernel. In this paper, the quantificational indexes of classification evaluation are discussed, when changing the number of classes, the spatial data autocorrelation degree is changed, this law is researched. Based on this law, the number of classes can be determined, and this new method is given in this paper. The result of data classification should be further evaluated by means of spatial distribution characteristic of data, and an example is given to test it.

Key words: statistical map; data classification; spatial autocorrelation; quality evaluation

About the first author: GUO Qingsheng, Ph. D, professor, Ph. D supervisor, engaged in the research on cartographic generalization, intelligent handling and visualization of geographical information.
E-mail: guoqingsheng@yahoo.com