DOI：10.13203/j.whugis20120562 文章编号：1671-8860(2014)04-0486-07

运用 Q 统计分析网络空间现象关联模式

田 晶[1] 何 遒[1] 周梦杰[1]

1 武汉大学资源与环境科学学院地理信息系统教育部重点实验室，湖北 武汉，430079

摘 要：空间关联模式的分析和发现是空间数据挖掘领域关注的热点问题。介绍了 Q 统计的概念，并将距离测度定义为网络最短路径。应用 Q 统计对一种网络空间现象——酒店的空间关联模式进行了分析。研究结果从空间数据挖掘的角度验证了酒店间的聚集效应。同时，也说明了 Q 统计在进行网络空间现象关联模式分析中是有效的。最后对欧式空间与网络空间下的 Q 统计进行了比较。

关键词：网络空间现象；Q 统计；空间关联；道路网；酒店

中图法分类号：P208 **文献标志码**：A

现实世界中，很多现象发生于网络或邻近网络，例如交通事故、基础设施点的分布等等。这些现象被称为网络空间现象，可抽象表达为网络上的点[1-2]。对于网络空间现象的分析一直是空间分析的热门研究问题。网络空间分析的显著特点是将距离测度由平面上的欧几里德距离变为网络上的最短路径距离。

空间关联是事物和现象在空间上的相互依赖、相互制约、相互影响和相互作用，是事物和现象本身固有的属性，是地理空间现象和空间过程的本质特征[3]。空间关联模式的分析和发现是空间数据挖掘领域关注的热点问题[4]，其研究符合地图制图学与地理信息工程学科发展的趋势[5]。其实现方法主要分为数据挖掘方法和空间统计方法[6-7]。而空间统计方法设计度量指标进行关联模式的分析。常用的指标或工具有 Moran's I、Geary's C、Q 统计等等[8-11]。本研究属于运用空间统计方法分析空间关联模式。

Ruiz 等[12-13]学者提出的 Q 统计是针对定性变量的分析而开发的统计量，其理论基础源于符号动力学。Q 统计能检测给定空间分布的定性变量的空间关联模式。Ruiz 等[12]将 Q 统计用于快餐店的关联模式分析；Paez 等[14]运用 Q 统计进行了城市人口聚群模式和曝光模式的分析，在这些研究中，距离测度均为平面欧几里德距离。

目前，对于网络空间现象的空间关联模式的分析研究较少，分析工具主要有网络交叉 K 函数，最近邻距离和条件最近邻距离[1,15]。本文引入 Q 统计来分析网络空间现象的关联模式。与已有的研究[12,14]不同，将 Q 统计中的距离测度定义为网络最短路径距离，以城市中的酒店这一类网络空间现象为对象，进行其关联模式分析，并与已有的酒店分布规律的研究结论进行对照。

1 网络空间下的 Q 统计

1.1 Q 统计的概念

Q 统计是针对定性变量的分析而开发的统计量，其理论基础源于符号动力学，能检测给定空间分布的定性变量的空间关联模式[12]。其基本原理是：将空间定性变量的分布作为离散的空间过程，借用符号动力学定义空间过程的符号，即得该空间过程的符号熵。该空间过程下的符号熵与独立(或随机)过程下的符号熵的似然比测试即为 Q 统计。根据统计学理论，与相应自由度的卡方分布的值进行比较，即可从全局判断该空间过程在统计意义下是否具有关联。其基本概念和定义的详细描述可参考文献[12]。由于本研究将距离测度定义为网络距离，所以称该 Q 统计是网络空间下的 Q 统计。

设分布于网络的定性变量 Y，其取值为 $\{a_1, a_2 \cdots a_k\}$，其分布的位置为 s_i，$i = 1, 2 \cdots N$，这些位置是固定的，则 $\{Ys_i\}$，$s_i \in s$ 是一个离散空间过程。如图 1 所示，分布在道路网的定性变量 Y，其

收稿日期：2013-03-22

项目来源：国家基础科学人才培养基金武汉大学地理科学理科基地资助项目(J1103409)。

第一作者：田晶，博士。主要研究方向为地图自动综合和空间数据挖掘。E-mail：yutaka-2010@163.com

取值为{black, white},分布的位置为 $s_i, i = 1,2\cdots11$。其中,定性变量取值数 k 为 2,black 出现了 5 次,其相对频率为 5/11,white 出现了 6 次,其相对频率为 6/11。

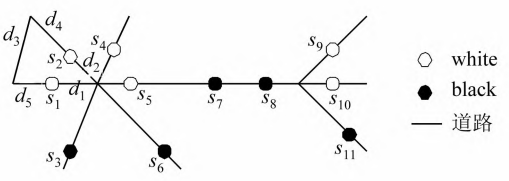

图 1　网络空间现象的分布

Fig. 1　Distribution of Network Spatial Phenomena

定义 1　网络空间现象的距离测度(distance measure of spatial network phenomena):分布在网络上的两点间的距离是两点沿着网络的最短路径,简称网络距离。

定义 2　m-环绕(m-surrounding):某一位置 S_i 的值及其按由近及远顺序排列的 $m-1$ 个最近的邻居的值构成的 m 维空间称为 m-环绕,其中 m 为 m-环绕的维数。对于网络空间,s_i 的 $m-1$ 个最近的邻居是与 S_i 网络距离最小的 $m-1$ 个位置。

定义 3　标准符号(standard symbols):由于一个 m-环绕包括 m 个位置,而每个位置可取 k 个值,所以一个 m-环绕具有 k^m 个可能的值。将每一个可能的值定义为一个标准符号,记作 $\sigma_j, j=1,\cdots, k^m$。记 $\Gamma=\{\sigma_1, \sigma_2,\cdots,\sigma_k^m\}$ 为所有的标准符号集合。

定义 4　观察样本(observations):定性变量 Y 出现的位置及其对应的值,其总数记为 N,图 1 中 N 为 11。

定义 5　符号化样本(symbolized observations):考虑所有观察样本,由 m-环绕的定义必然出现重叠,这无法满足统计测试中的一些假设条件[12, 14],为了控制重叠,需要选择观察样本中的一部分位置,然后进行符号化,符号化样本数记为 R。

定义 6　重叠度(overlapping degree):对于任意两个位置 s_i 和 s_j,s_i 的 m-环绕与 s_j 的 m-环绕的最大重叠数,记为 r,且 $0\leqslant r<m$。

m-环绕的维数 m,观察样本数 N,符号化样本数的最大值 $R_{\max}$ 以及重叠度 r 的关系为[12]:

$$R_{\max}=\left[\frac{N-m}{m-r}\right]+1 \tag{1}$$

式中,"[]"为取整。

定义 7　符号的绝对频率(absolute frequency of symbols):符号 σ_j 的绝对频率即为分布位置中符号为 σ_j 类型的位置总数,记作 $n\sigma_j$。

定义 8　符号的相对频率(relative frequency of symbols):符号 σ_j 的相对频率为该符号的绝对频率与符号化样本数之商,记作 $p\sigma_j$。

定义 9　符号熵(symbolic entropy):空间过程中包含的信息量,记作 $h(m)$,计算公式如下:

$$h(m)=-\sum_{\sigma_j\in\Gamma}p\sigma_j\ln(p\sigma_j) \tag{2}$$

当某一符号占据主导地位,则符号熵 $h(m)$ 趋近于 0;若定性变量取值的相对频率相等,且每个符号出现的相对频率相等,则符号熵有最大值 $\ln(k^m)$,所以 $0<h(m)\leqslant\ln(k^m)$。

定义 10　Q 统计(Q statistic):被观察的空间过程的符号熵与独立(随机)过程下的符号熵的似然比测试。独立过程下,若定性变量取值的相对频率相等,且每个符号出现的相对频率相等,符号熵的上界为 $\ln(k^m)$,然而实际中,由于定性变量取值的相对频率不等,如图 1 中 black 为 5/11,white 为 6/11,所以独立过程中,在定性变量取值相对频率不等的情况下符号熵的上界 η 为:

$$\eta=\sum_{i=1}^{k^m}\frac{n\sigma_i}{R}\sum_{j=1}^{k}\alpha_{ij}\ln(q_i) \tag{3}$$

式中,α_{ij} 为值 α_j 在符号 σ_i 中出现的次数;q_j 为值 α_j 出现的相对频率。

则 Q 统计为:

$$Q(m)=2R(\eta-h(m)) \tag{4}$$

如果空间过程是独立(随机)的,则 Q 统计服从近似的自由度为 k^m-1 的卡方分布 $\chi^2_{k^m-1}$,其证明过程详见文献[12]的附录。

设 $0\leqslant\beta\leqslant1$,置信水平 $100(1-\beta)\%$ 下拒绝空间独立的决策规则为:如果 $Q(m)>\chi^2_{k^m-1\beta}$,则拒绝空间独立的假设,否则不拒绝空间独立的假设。

定义 11　等价符号(equivalent symbols):仅考虑 m-环绕中定性变量值出现的次数,不考虑出现的次序,所构成的符号称为等价符号,其可能出现的符号总数为 $k(k+1)\cdots(k+m-1)/m!$。与标准符号相比,等价符号没有了拓扑信息,但是大大减少了符号的数量,这对于实际应用具有重要意义,因为随着 m 和 k 的增加,必然导致符号爆炸式地增加。

1.2　参数设定原则

由 §1.1 的描述可知,为了计算 Q 统计,需要确定一系列的参数。这些参数包括观察样本数 N,符号化样本数 R,定性变量类别数 k,m-环绕中的 m 和 r,其他诸如符号的相对频率,定性变量类别的相对频率都可以从上述参数中导出。

观察样本数 N,定性变量类别数 k 由具体的问题和相应的实验决定,可以说是确定的。而 m-环绕中的尺寸 m 和 r 是不确定的,需要根据实际情况设定。这里给出参数设定的指导性原则。根据文献[12]设定的规则,由于实际中用到了卡方分布,所以要求符号化样本数 $R\geqslant 5k^m$。

由于 $N>R\geqslant 5k^m$,则可知 $m<\ln(N/5)/\ln k$,所以,m 的取值范围是 $2\leqslant m<\ln(N/5)/\ln k$,且 m 为整数。最后,可结合实际情况确定 m 的值。

当 m 确定后,可以确定重叠度 r,由于符号化样本数最大值 $R_{\max}=[(N-m)/(m-r)]+1$,其中"[]"是取整,且 $R\geqslant 5k^m$,所以 r 也可确定。重叠度 r 越大,则保留的观察样本越多,这样会增加Q统计的效力,但同时会增加第一类错误,即拒绝实际上为空间独立的假设的风险[12]。

2 案例分析

酒店是一类备受关注的网络空间现象。酒店的区位选择对于酒店的建立与发展至关重要。影响酒店分布的因素通常有:星级、配套实施、房间数、所有权、一定距离内其他类型的酒店、交通便利性、与旅游景点的距离、与商业服务的距离等等[16]。其中一定距离内其他类型的酒店是研究的一个重点问题,解释的理论主要有地理学理论、经济学理论及市场理论[17-19]。Kalnins 和 Chung[17]提出了酒店分布的聚集效应原理,他们提出一系列的定性假设,并运用条件 logit 的方法对假设进行了验证。该研究从经济学和市场学的角度回答了在酒店选址时,是否独立于周围已有的酒店。他们的研究表明高档酒店更倾向于进入一个同样是高档酒店存在的市场,低档酒店倾向于进入高档酒店存在的市场。Yang 等人[16]关于酒店区位选择的 logit 模型得到了类似的结论。

从空间关联的角度,上述问题转化为:哪几种类型的酒店频繁地分布在一起,同类型酒店的分布趋于排斥还是相互吸引。本研究运用网络空间的Q统计进行酒店空间关联模式分析,并与已有的研究结论进行对照。

2.1 数 据

深圳市酒店数据由深圳市基础地理信息中心提供,为2009年的POI数据子集。

在此,定性变量即为分布在道路网周围的不同类型的酒店。不同的国家采用不同的酒店分类标准[16,20]。本研究参照文献[16]的分类方法,将深圳市酒店分为3类:低档酒店(budget hotel,B),包含0-2星酒店1 503家;中档酒店(middle hotel,M),包括3星酒店222家;高档酒店(luxery hotel,L),包含4～5星级酒店186家。其分布如图2所示。

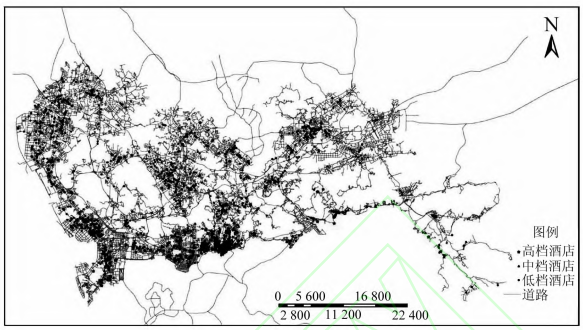

图2 深圳市酒店分布

Fig. 2 Distribution of Hotels in Shenzhen City

2.2 结果与分析

根据§1.2的描述,首先要进行参数的选择和配置。定性变量的类别数 $k=3$,m 和 r 不确定,下面详述它们在本例中的设定方式。

本例中,$N=1\ 911$,$k=3$,由 $\ln(N/5)/\ln(k)=5.412\ 2$,可知 m-环绕的维数 m 可以为5、4、3、2。那么应该怎么选择呢?在已有的酒店分布规律的研究中,一定距离内其他类型的酒店是一个考虑因素,而这个一定距离有的学者定义为4～5 km[12],亦即常提到的方圆4、5 km,这里的距离是指欧几里德距离。考察每一个样本点到其4阶近邻(对应于5-环绕)的距离分布,如图3所示,超过4 km的仅占1.58%,超过5 km的仅占1.1%,而且此处定义的距离是网络距离,两点的网络距离要大于等于两点的欧几里德距离。由此可见,m-环绕的维数为5,即考虑4阶邻居是合理的。

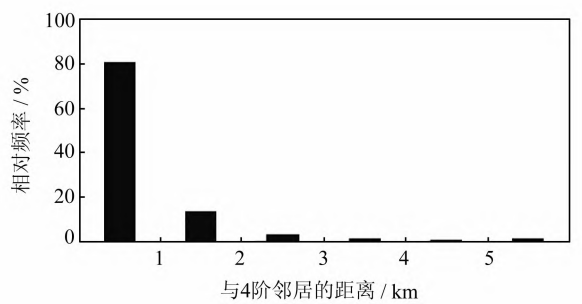

图3 酒店与其4阶邻居的距离分布

Fig. 3 Distribution of Distance Between a Hotel and Its 4th Order Neighbour

由 $m=5$,$k=3$ 可得,最少需要的符号化样本数为 $5k^m=1\ 215$,重叠度的取值为1、2、3、4时得出的符号化样本数最大值分别为477、636、953和1 907,只有重叠度为4时,满足要求 $R\geqslant 5k^m$,所

以本例中重叠度 r 取值为 4。基于此,参数的设定和计算结果如表 1。

表 1　参数与计算结果

Tab.1　Summary of Parameters and Result

观察样本数 N	1 911		
符号化样本数 R	1 907		
定性变量取值数 k	3		
m-环绕的维数 m	5		
重叠度 r	4		
标准符号数 k^m	243		
等价符号数 $k(k+1)\cdots(k+m-1)/m!$	21		
定性变量值出现的相对频率(q_i)	高档酒店 0.097 3	中档酒店 0.116 2	低档酒店 0.786 5
空间关联测试	Q 值	卡方分布自由度	P 值
标准符号	488.573 4	242	0.000 0
等价符号	5 734.730 4	20	0.0000

自由度 $3^5-1=242$ 的卡方分布,0.05 显著性水平的临界值为 279.287 6,相应 Q 值为 488.573 4,自由度 $3^5-1=20$ 的卡方分布,0.05 显著性水平的临界值为 31.410 4,相应 Q 值为 5 734.730 4,说明应拒绝空间独立的假设,表明酒店在分布上不是空间独立的。既然分布并非独立的,就必然存在一定的分布模式。下面对此进行必要的分析。

由于等价符号与关联模式挖掘中的项集具有相似性,而且现在仅关心哪些酒店频繁地出现在一起,所以应用等价符号进行分析是合理的。本研究中的等价符号等同于 5 项集。它们在独立状态下其出现的相对频率是 1/21,在本例中,符号化样本数为 1 907,则每个符号近似出现 91 次。

那么,实际这些等价符号出现了多少呢? 哪些等价符号为出现的较为频繁? 亦即哪些是频繁项集。

在本例中,出现的等价符号 18 个。虽然有些符号出现的相对频率较高,但是并不是统计意义上的高,将其在独立过程下的相对频率期望值及其置信区间来表示,如图 4 所示。

图 4 柱状图中的每一个柱的高度表示实际情况下每个等价符号出现的相对频率,点圈线表示独立过程下该符号出现的相对频率期望值及其 95%的置信区间,置信区间的计算方法详见文献[14]的附录。图 4 中浅色填充的柱表示显著高于独立过程下期望相对频率的模式,深色填充的柱表示显著低于独立过程下期望相对频率的模式,无填充的柱表示不显著高于或低于独立过程下期望相对频率的模式。

10 种等价符号显著高于独立过程下的期望

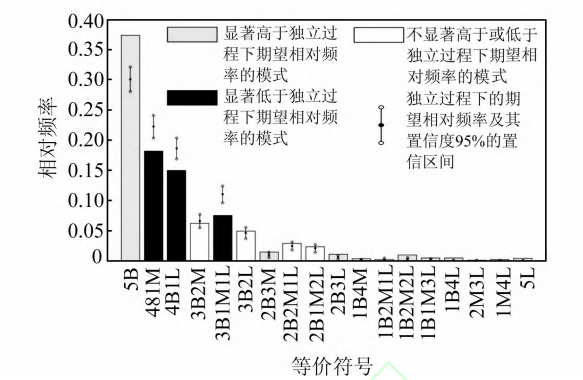

图 4　实际情况的等价符号相对频率和在独立过程下的期望相对频率

Fig.4　Relative Frequency of Equivalent Symbols Under Practical and Independent Process

相对频率,包括 5B、2B3M、2B3L、1B4M、1B2M2L、1B1M3L、1B4L、2M3L、1M4L 和 5L。5B,5L 表明高档酒店和高档酒店聚集,低档酒店与低档酒店聚集,反映了异质市场中的寻求资源的聚集效应。聚集效应是指产业内或与产业有关的生产要素在特定地理范围的高度集中的现象[21],这种聚集是为了使有限资源的价值最大化,提高效益。2B3M、2B3L、1B4M、1B2M2L、1B1M3L、1B4L 表明低档酒店倾向于同中高档酒店毗邻,以获取中高档酒店的溢出价值。这种溢出价值包括以下几个方面:首先是劳动市场的共享,高档酒店一般规模较为庞大,人员素质较高,流通大,这为其周边酒店提供了劳动力市场;其次是附属产业的增长,每一个附属产业虽然只服务于生产过程中一个很小的分支,但它为附近的许多产业工作降低了使用费用[22];最后是知识技术的外溢,比如经营管理的理念方法等[21]。然而,少量中高档酒店的溢出价值的确无法满足低档酒店,三种符号显著低于独立过程下的期望相对频率,包括 4B1M、4B1L、3B1M1L,说明中高档酒店的溢出价值有限,不可能同时满足很多低档酒店的需要。同样地,两种等价符号 3M2L、4M 1L 未出现,少量高档酒店不与中档酒店毗邻,说明少量高档酒店的溢出价值同样不能满足中档酒店的需要。还有 1 种等价符号 5M 没有出现,原因可能是中档酒店之间的竞争大于它们之间共同寻求资源的需求。由本研究得出的结果从空间数据挖掘的角度验证了 Kalnins 和 Chung[17] 的关于在异质市场中酒店区位选择的结论,说明 Q 统计在进行网络现象关联模式的分析上是有效的。

2.3 比 较

将酒店间的距离定义为欧式距离，运用§2.1数据进行了对比实验，出现的等价符号如图5所示，与图4相比共有三种等价符号发生了显著变化。首先，1B4M、1B1M3L两种等价符号由显著高于独立过程下期望相对频率的模式变为不显著高于或低于独立过程下期望相对频率的模式；其次，3B2L由不显著高于或低于独立过程下期望相对频率的模式变为显著高于独立过程下期望相对频率的模式。

在网络空间下，1B4M和1B1M3L出现频率显著高于独立过程下期望相对频率的模式说明低档酒店有靠近中高档酒店的倾向，并分享其溢出价值，与已有研究结论相似[17]。3B2L不显著高于或低于独立过程下期望相对频率的模式同样印证了已有研究中关于中高档酒店的溢出价值有限，不可能同时满足很多低档酒店的需要的论述[17]。而在欧式空间下，1B4M和1B1M3L出现频率不再显著高于独立过程下期望相对频率的模式使低档酒店倾向于同中高档酒店毗邻，以获取中高档酒店的溢出价值的结论支持减弱，同时3B2L的变化让低档酒店的过度聚集显得更为突出，这对于分享溢出价值是不利的。综上所述，本研究认为网络空间下得出的结论更为合理。

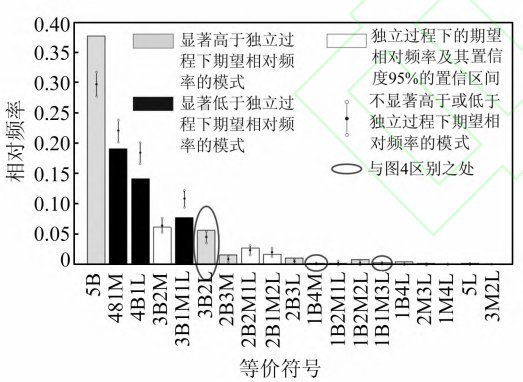

图5 欧式空间下实际情况的等价符号相对频率和在独立过程下的期望相对频率

Fig. 5 Relative Frequency of Equivalent Symbols Under Practical and Independent Process in Euclidean Space

3 结 语

本文介绍了Q统计的概念，将其距离测度定义为网络上的最短路径距离，扩展为网络空间下的Q统计。对酒店的空间关联模式进行了分析，并运用经济学和市场学的研究结论解释了本文的结果。下一步的研究主要在两个方面展开：第一，与网络交叉K函数、最近邻距离和条件最近邻距离进行比较研究；第二，本文在进行分析时仅考虑了酒店与酒店间的关联模式，在酒店区位选择时，交通便利、周边服务也是考虑到重要因素，酒店是否和地铁站这一类交通设施存在关联也将是下一步需要深入研究的问题。

参 考 文 献

[1] Okabe A, Yamada I. The K-function Method on a Network and Its Computational Implementation[J]. *Geographical Analysis*, 2001, 33(3): 271-290

[2] Okabe A, Okunuki K, Shiode S. SANET: A Toolbox for Spatial Analysis on a Network[J]. *Geographical Analysis*, 2006, 38(1): 57-66

[3] Ma Ronghua, Pu Yingxia, Ma Xiaodong. Mining Spatial Association Patterns from GIS Database [M]. Beijing: Science Press, 2007(马荣华，浦英霞，马晓冬. GIS空间关联模式发现[M]. 北京：科学出版社，2007)

[4] Li Deren, Li Deyi, Wang Shuliang. Spatial Data Mining Theories and Applications[J]. *Geomatics and Information Scicence of Wuhan University*, 2001, 26(6): 492-499(李德仁，李德毅，王树良. 论空间数据挖掘和知识发现[J]. 武汉大学学报·信息科学版，2001，26(6)：492-499)

[5] Wang Jiayao. Development Trends of Cartography and Geographic Information Engineering[J]. *Acta Geodaetica et Cartographica Sinica*, 2010, 39(2): 115-119(王家耀. 地图制图学与地理信息工程学科发展趋势[J]. 测绘学报，2010，39(2)：115-119)

[6] Huang Yan, Shekhar S, Xiong Hui. Discovering Colocation Patterns from Spatial Data Sets: A General Approach[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2004, 16(12): 1 472-1 485

[7] Bembenik R, Rybinski H. FARICS: A Method of Mining Spatial Association Rules and Collocations Using Clustering and Delaunay Diagrams[J]. *Journal of Intelligent Information Systems*, 2009, 33(1): 41-64

[8] Ripley B D. The Second-order Analysis of Stationary Point Processes[J]. *Journal of Applied Probability*, 1976, 13(2): 255-266

[9] Ord J K, Getis A. Local Spatial Autocorrelation Statistics: Distributional Issues and Application[J]. *Geographical Analysis*, 1995, 27(4): 286-306

[10] Guo Luo, Du Shihong, Haining R, et al. Global and Local Indicator of Spatial Association Between Points and Polygons: A Study of and Use Change [J]. *International Journal of Applied Earth Ob-*

servation and Geoinformation, 2013,21:384-396

[11] Hu Wei. Co-location Pattern Discovery[M]//Shash S, Hu X. Encyclopedia of GIS. Berlin: Springer-Verlag, 2008

[12] Ruiz M, Lopez F, Paez A. Testing for Spatial Association of Qualitative Data Using Symbolic Dynamics[J]. *Journal of Geographical Systems*, 2010, 12(3): 281-309

[13] Ruiz M, Lopez F, Paez A. Comparison of Thematic Maps Using Symbolic Entropy[J]. *International Journal of Geographical Information Science*, 2011, 26(3): 413-439

[14] Paez A, Ruiz M, Lopez F, et al. Measuring Ethnic Clustering and Exposure with the Q Statistic: An Exploratory Analysis of Irish, Germans, and Yankees in 1880 Newark[J]. *Annals of the Association of American Geographers*, 2010, 102(1): 84-102

[15] Spooner P, Lunt I D, Okabe A, et al. Spatial Analysis of Roadside Acacia Population on a Road Network Using the Network *K*-function[J]. *Landscape Ecology*, 2004, 19(5): 491-499

[16] Yang Yang, Wong K, Wang Tongkun. How do Hotels Choose Their Location? Evidence from Hotels in Beijing[J]. *International Journal of Hospitality Management*, 2012, 31(3): 675-685

[17] Kalnins A, Chung W. Resource-seeking Agglomeration: a Study of Market Entry in the Lodging Industry[J]. *Strategic Management Journal*, 2004, 25(7): 689-699

[18] Chung W, Kalnins A. Agglomeration Effects and Performance: a Test of the Texas Lodging Industry [J]. *Strategic Management Journal*, 2001, 22 (10): 969-988

[19] Egan D J, Nield K. Towards a Theory of Intraurban Hotel Location[J]. *Urban Studies*, 2001, 37 (3): 611-621

[20] Fernandez M, Bedia A. Is the Hotel Classification System a Good Indicator of Hotel Quality? An Application in Spain[J]. *Tourism Management*, 2004, 25(6): 771-775

[21] Xu Kangning. The Source of Industry Agglomeration[M]. Beijing: People's Press, 2006 (徐康宁. 产业聚集形成的源泉[M]. 北京: 人民出版社, 2006)

[22] Marshall A. Principles of Economics[M]. 8th Edition, London: Macmillan and Co. Ltd, 1961

Spatial Association Analysis of Network Spatial Phenomena with the Q Statistic

TIAN Jing[1]　*HE Qiu*[1]　*ZHOU Mengjie*[1]

1　Key Laboratory of Geographic Information System, School of Resource and Environment Science, Wuhan University, Wuhan 430079, China

Abstract: The analysis and discovery of spatial association is a hot issue in the field of spatial data mining. However, a little attention has been paid to the spatial association of network spatial phenomena. The objective of this article is to demonstrate the application of the Q statistic, developed for the analysis of the spatial association of qualitative variables, to the detection of spatial association of the network spatial phenomena. This paper introduces the Q statistic concept, and defines distance measure by the shortest path. The spatial association of the hotels in Shenzhen city in China are analyzed using the Q statistic. A comparative analysis of the Q statistic in Euclidean space and network space was conducted. The results show the agglomeration effects among hotels from a spatial data mining perspective. Results also show that the Q statistic is valid for spatial association analysis of network spatial phenomena.

Key words: network spatial phenomena; Q statistic; spatial association; road network; hotel

First author: TIAN Jing, PhD, specializes in automated map generalization and spatial data mining. E-mail: yutaka-2010@163.com

Foundation support: The Project for National Basic Science Personnel Training Fund, No. J1103409.