·监测·

空间流行病学分析中细小单元的区域化归并方法：以义乌市甲状腺癌为例

滕世助　贾巧娟　黄以坚　陈亮操　费徐峰　吴嘉平

【摘要】 **目的**　空间流行病学研究中，细小地域单元由于人口数目少，偶发的个别病例出现就有可能导致极值的产生，给真实发病的分析带来困难。本研究以浙江省义乌市甲状腺癌2010—2013年监测数据和人口资料为基础，利用不同的区域划分描述甲状腺癌空间分布规律，探测病例空间聚集热点，为进一步开展环境和人群监测提供参考依据，同时评价小单元区域化在空间流行病学研究中的应用价值。**方法**　采用基于地理环境相似性（同质性检验）和空间相邻原则集群和分区的区域化方法，利用地理信息系统（GIS）将人口较少村落合并为具有足够人口基数的地理单元，以增加率值估计的稳定性和可靠性。**结果**　研究区新产生的区域单元具有足够的人口基数，产生的甲状腺癌发病率稳定和正态化。对新的区域单元进行热点探测（Getis-Ord）发现研究区中部平原地区有显著的高发病聚集。**结论**　小单元区域化方法能有效解决人口基数过小问题，有助于空间热点的探测与分析，能为进一步探寻甲状腺癌的危险因素提供依据。

【关键词】 地理区划；甲状腺癌；空间分析

Hierarchical regionalization for spatial epidemiology: a case study of thyroid cancer incidence in Yiwu, Zhejiang

Teng Shizhu[1], *Jia Qiaojuan*[1], *Huang Yijian*[1], *Chen Liangcao*[1], *Fei Xufeng*[2], *WU Jiaping*[2]. 1 *Yiwu Center for Disease Control and Prevention*, *Yiwu* 322000, China; 2 *College of Environmental and Resource Sciences*, *Zhejiang University*

Corresponding author: *Jia Qiaojuan*, Email: 605515026@qq.com

This work was supported by a grant from the Health Department of Zhejiang Province (No. 2014KYB299).

【**Abstract**】 **Objective**　Sporadic cases occurring in mall geographic unit could lead to extreme value of incidence due to the small population bases, which would influence the analysis of actual incidence. **Methods**　This study introduced a method of hierarchy clustering and partitioning regionalization, which integrates areas with small population into larger areas with enough population by using Geographic Information System (GIS) based on the principles of spatial continuity and geographical similarity (homogeneity test). This method was applied in spatial epidemiology by using a data set of thyroid cancer incidence in Yiwu, Zhejiang province, between 2010 and 2013. **Results**　Thyroid cancer incidence data were more reliable and stable in the new regionalized areas. Hotspot analysis (Getis-Ord) on the incidence in new areas indicated that there was obvious case clustering in the central area of Yiwu. **Conclusion**　This method can effectively solve the problem of small population base in small geographic units in spatial epidemiological analysis of thyroid cancer incidence and can be used for other diseases and in other areas.

【**Key words**】 Geographic regionalization; Thyroid cancer; Spatial analysis

癌症是一种低发病率的疾病，小尺度癌症数据分析面临因人口基数过小而产生的发病率估计不稳定、对缺失数据敏感性等问题[1]。近年来，地理信息系统（GIS）在空间数据分析、处理和模型模拟方面的功能被广泛地应用于癌症研究[2]。随着GIS技术的发展和癌症监测数据的标准化，根据患者详细的住址信息，可以精确地将其定位到镇（街道）、村（社区）甚至是经纬度坐标，使得从更加精细的空间尺度研究癌症的聚集性成为可能[1]。但是，高精度的空间尺度必然造成研究单元的人口基数过小，导致不可靠的空间格局评估和不正确的发病聚集模式[1,3]。人口基数过小而产生的极值问题是不可避免的，不稳定极值的存在将影响空间格局分析的正确性。另

DOI: 10.3760/cma.j.issn.0254-6450.2015.10.023
基金项目：浙江省医药卫生科技计划项目（2014KYB299）
作者单位：322000 浙江省义乌市疾病预防控制中心（滕世助、贾巧娟、黄以坚、陈亮操）；浙江大学环境与资源学院（费徐峰、吴嘉平）
通信作者：贾巧娟，Email：605515026@qq.com

外，虽然镇（街道）尺度拥有足够的人口基数，能够得到相对稳定的甲状腺癌发病率分布情况，但其无法精确表达同一镇（街道）内甲状腺癌发病率的空间变异情况。因此，需要一种合适的区域化方法来获得能够兼顾稳定性和正确性的空间表达尺度[3]。研究人员已提出许多区域化的方法，比如利用GIS和皮亚诺曲线对地理单元进行排序合并[4]；在GIS系统中利用最邻近法将单元合并[5]；随机区域化后，通过反复迭代重新分配合并单元等方法[6]，但这类方法未考虑研究区内自然环境的异质性，可能会将环境差异显著的单位合并，不利于癌症空间格局与环境分布的关联研究。本文采用一种基于地理环境相似性和空间连续性区域化方法[7]，将义乌市村镇合并为具有一定人口基数和相似地理环境的单元，用以表达当地甲状腺癌发病率的空间格局，为进一步开展相关环境风险因子的研究提供基本参考和科学基础。

资料与方法

1. 数据来源：义乌市位于浙江省中部，东、南、北三面群山环抱，为丘陵地区，境内有中低山、丘陵、岗地、平原，土壤类型多样，光热资源丰富，属亚热带季风气候，温和湿润，四季分明。该市为商贸名城，小商品批发集散地享誉国内外，服装、饰品、针织等加工为主的工业企业发达，人均收入水平位居全省前列。该市从2009年3月开始采用国际癌症登记协会（IACR）推荐的单机录入软件CanReg 4，统一对癌症数据进行管理。每年定期从浙江省慢性病监测信息管理系统中导出肿瘤发病和死亡数据，进行整理、审核、漏报调查和统计。根据义乌市统计局2013年统计年鉴，义乌市2012年底共有794个行政村，户籍人口数为753 314人。本次研究采用2010—2013年该市全部确诊的甲状腺癌患者558例，其中男性152例（占27.2%），女性406例（占72.8%），总体粗发病率为18.36/10万。

2. 地理数据来源及预处理：研究所采用的义乌市行政区划图为浙江省测绘局提供的行政村区划图。该市的高程分布提取于30 m空间分辨率的ASTER GDEM数据（http://datamirror. csdb.cn/index.jsp/），土地利用信息来自于对2009年Landsat ™影像的目视解译[8]。所有图像采用Universal Transverse Mercator投影和World Geodetic System 84坐标系。利用地理信息系统定位和编码将558例甲状腺癌病例分配到具体的行政村，根据794个行政村的发病人数及人口基数计算4年平均发病率，其分布情况见图1（村尺度）。同时计算各个村的平均海拔，平均坡度，建筑、林地和农田用地面积比例。考虑到以上5个地理属性的共线性问题，本研究通过因子分析[9]，将5个变量简化成2个主成分（见表1），可以解释原始变量的77.5%变化。对2个主成分进行正态转换后用于区域化研究中地理环境的同质性检测。

表1 因子成分表

因　素	成份1	成份2
平均海拔	0.791	0.347
平均坡度	0.818	0.300
建设用地面积（%）	−0.814	0.552
林地面积（%）	0.889	−0.309
农田面积（%）	−0.316	−0.648

3. 小单元区域化研究：义乌市2010—2013年间分村甲状腺癌发病率的分布见图1（村尺度）。4年间义乌市甲状腺癌发病率高于100.00/10万的有21个行政村（红色），发病较高的行政村80.95%人口基数＜1 000人，而发病人数＜5人。人口基数过小对义乌市甲状腺癌发病率聚集性分析的影响很大，极值地区的甲状腺癌发病率不适于统计分析，同时对空间格局分析也存在巨大的干扰。为解决人口基数过小的问题，需要进行区域化处理。区域化的目的是将属性相似（指海拔等地理属性）、连续（具有共同边界）、人口基数较小的村落合并为具有足够人口基数的新区域以保证发病率估计的稳定性。本次研究采用基于动态约束凝聚集群和分区的区域化方法（REDCAP）[3]对行政村进行集群和分区。本质上REDCAP的目标是通过聚集相邻且具有相似属性值（如地理信息）的小区块，构建一系列同质性的区域。为了达到这一目的，REDCAP构建了以小区块之间属性相似性为基础的聚类分层，并划分空间上相邻的集群来精确地优化同质性检验。同质性检验采用所有属性的总方差来（TSD）评价，TSD越小代表同质性越高[10]，方程如下：

$$TSD = \sum_{r=1}^{k}\sum_{i=1}^{n_r}\sum_{j=1}^{d}(x_{ij} - \bar{x}_j)^2 \quad (1)$$

其中k是新的地理区域的个数，n_r是地理区域r中基本行政村的个数，d是用于区划地理属性的个数，x_{ij}是具体的变量值，$\bar{x}_j$是地理区域内变量j的平均值。

如图2所示，REDCAP由两个步骤组成：①连续约束的层次聚类；②自上而下的分区。每一个多边形的颜色代表其属性值，相似的颜色表明相近的

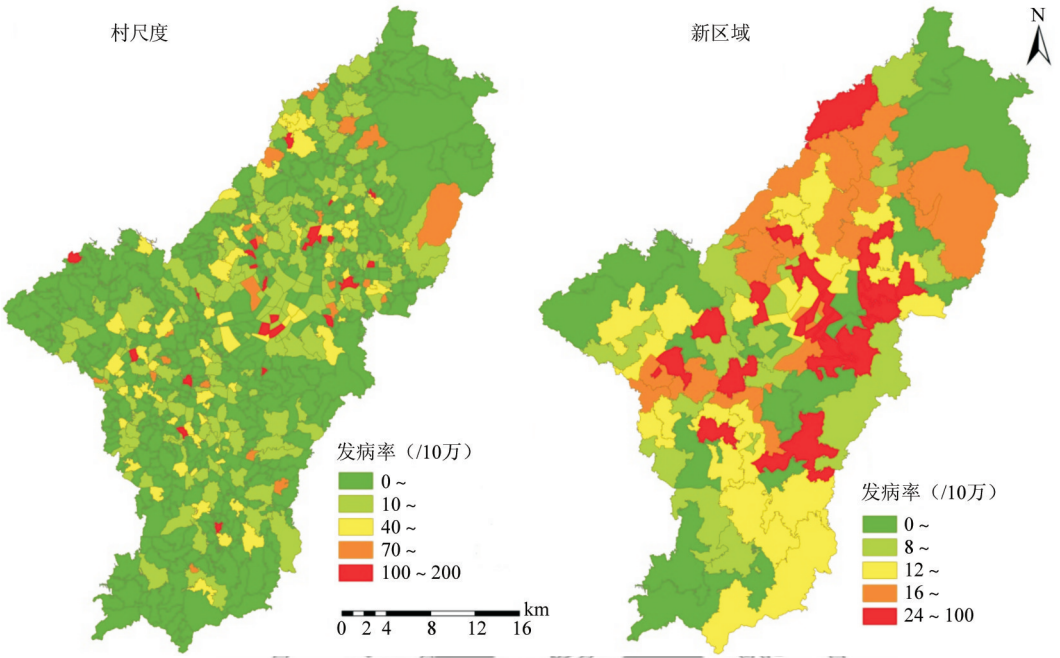
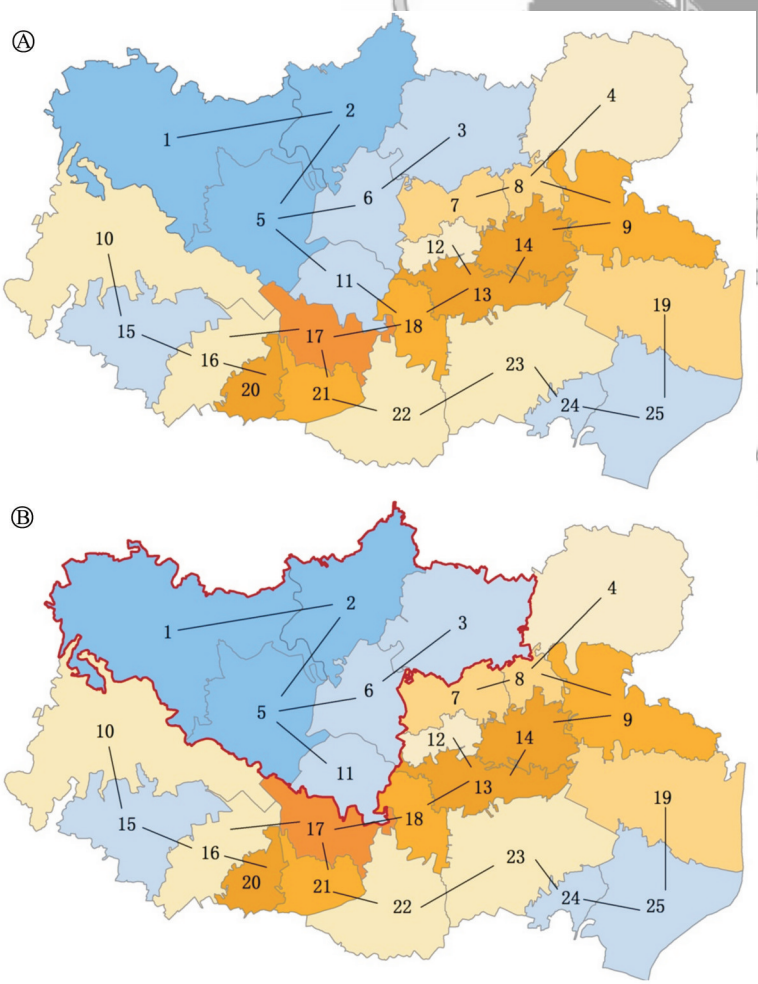


图1　村尺度/新区域甲状腺癌发病率分布



注：Ⓐ连续约束的层次聚类；Ⓑ自上而下的分区

图2　区域化方法示例

值。如果两个多边形存在共同边界，则认为两者相邻。在第一步中，如图2Ⓐ所示，REDCAP构建了在连续约束并基于属性相似度的空间相邻集群的分层。两个相邻且最相似的区块将融合形成更高级别的集群，直到整个研究区域成为一个集群。在第二步中，如图2Ⓑ所示，REDCAP划分该集群并通过移除满足同质性的最优边界（如图2Ⓑ中的11 ~ 18）以产生两个区域。换句话说，这两个区域是根据每个区域内的同质性最大来确定产生的。划分不断进行直至满足期望的区域数量。对于癌症数据分析而言，特别是针对缓解人口数量较少的问题，REDCAP进行了改进，提供了额外的约束条件进行加强，例如区域人口最小数量。这些约束条件在第二步（如树状图的划分）中得到执行。对于每个潜在的分割，如果不能产生两个区域均满足这些约束条件，则该分割不予考虑。最终得到的区域均足够大并且每个区域都具有最高的同质性。本研究中考虑到要产生稳定的发病率数据最小需要的人口基数为20 000人[3]，研究的数据是义乌市4年平均的甲状腺癌发病率数据，所以采用的限制条件是人口基数＞5 000人（20 000/4＝5 000）。在GIS

中根据行政村所属的地理区域信息，统计各个区域4年内甲状腺癌的发病人数和人口基数，计算其发病率如图1（新区域）所示。

4. 空间聚集性探测方法：Getis-Ord统计模型被用于行政村和划分的地理区域甲状腺癌发病率的空间热点探测[11]。通过计算z得分的方法来检测空间上显著的甲状腺癌高发/低发聚集区。z得分的计算方程：

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2\right]}{n-1}}} \quad (2)$$

其中G_i^*是z得分，x_j是区域j的甲状腺癌发病率，$w_{i,j}$是邻域i相对于区域j的权重，本研究采用反距离加权的方法[12]。n是研究区所有区域的个数。$\bar{X}$和S的计算公式：

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (3)$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2} \quad (4)$$

z得分的绝对值＞2.58、2.58～1.96、1.96～1.65分别代表显著性在0.01、0.05、0.10水平的聚集分布，正值为高发聚集，负值为低发聚集。z得分处于−1.65～1.65则代表发病率的随机分布。

5. 统计学分析：本研究采用REDCAP软件（http://www.spatialdatamining.org/software/redcap）对研究区进行地理区划。运用Arcgis 9.3软件进行属性分析和提取、制作甲状腺癌发病分布图和空间热点探测。主成分分析及其他常规统计分析在SPSS 19.0软件中完成。

结　果

从2010年1月1日到2013年12月31日，检测出558例甲状腺恶性肿瘤患者，4年的病例数分别为92、134、136、196例，女性约为男性患者的3倍。期间，义乌市甲状腺恶性肿瘤发病率年均增长28.67%（从12.21/10万增加到26.02/10万）。男性发病率从4.88/10万增加到13.02/10万，年变化率约为38.67%；女性发病率从20.42/10万增加到40.84/10万，年变化率为25.99%。

义乌市794个村落中，有786个村的人口基数＜5 000，为保证甲状腺癌发病率估计的可靠性，约98.99%的村落需合并以达到5 000人的人口下限。在进行区域化后，共产生96个新的区域单元。原始村落和新区域的发病及人口描述性统计见表2。从人口基数上来看原始村落最小的人口基数只有53人，而进行合并后区域的人口基数均达到5 000人以上（最小5 002人），保证了甲状腺癌发病率估计的可靠性。同时，新区域所估计的发病率极值降低了近50%（新区域100.10/10万 *vs.* 原始区域199.60/10万），新区域的平均发病率也更接近全市的总体平均发病率，发病率的估计具有较高的稳定性：新区域甲状腺癌发病率的标准差为14.10，远小于原始村落发病率的标准差（29.14）；从图3可以看出，原始村尺度上有516个村的发病率为0，频率分布存在严重的左倾现象，而新区域上发病率分布更加趋向于正态分布，其峰值接近于总体的平均发病率。

表2 村尺度/新区域甲状腺癌发病描述性统计

因　素	村落	新区域
单元数	794	96
最大人口数	11 155	16 813
最小人口数	53	5 002
平均人口数	921	7 614
人口数标准差	943	2 373
最大发病率	199.60	100.10
最小发病率	0	0
平均发病率	15.51	17.66
发病率标准差	29.14	14.10

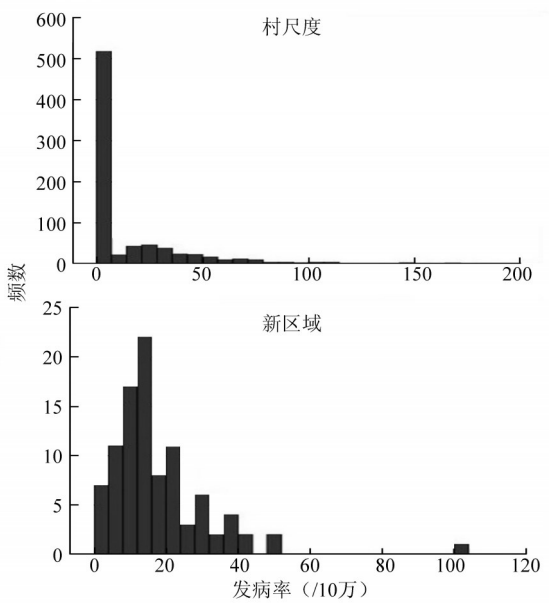


图3 村尺度/新区域甲状腺癌发病率频率分布

从图1（村尺度）中可以看出原始村落的甲状腺癌发病率分布十分散乱，发病率变动大。Getis-Ord统计模型空间热点探测结果也提示，原始村落高发

病聚集区呈零星分布的模式，没有明确的高发病聚集区域（图4，村尺度）[11]。进行区域化后的甲状腺癌发病率分布比较集中（图1，新区域），高发病率集中在中部平原的“心脏”地区（图4，新区域）。进行区域化后的发病率较稳定、可靠，更适合于进行热点探测等空间统计分析。

讨　论

对于癌症数据在行政村尺度上的空间聚集性问题，戚晓鹏等[2]对研究区内4种消化道癌症的死亡监测数据进行了空间自相关和空间热点探测分析，发现癌症死亡率在4 300 m尺度存在显著空间正相关，并探测到3个死亡率的高发地区。葛辉等[13]分析2005—2010年灵璧县食道癌死亡率的空间分布特征，发现该县中部特别是中东区域存在显著的高值聚集区。但是这些研究未对村落进行相应的多维度融合，没有妥善解决村落人口基数过小的问题，产生的死亡率估计和空间分布规律不够准确。

对癌症数据进行行政村尺度的空间聚集性分析不可避免地会遇到人口基数过小的问题，从而造成发病率估算的不稳定。在人口稀少的地区产生发病率的极高值，对发病的空间聚集性分析产生干扰。本研究采用基于地理环境相似性和空间连续性的方法，通过各个行政村的海拔、坡度和土地利用信息将村落合并为具有一定人口基数的较大区域。

分析发现，原始村尺度上发病率呈现极不稳定频率分布：516个村的发病率为0，其最大值（199.60/10万）与最小值（0）相差极大，平均值（15.54/10万）远小于实际平均发病率（18.36/10万），存在严重的左倾现象。而在新建村区域中，仅有7个区域的发病率为0，平均值（17.80/10万）更接近总体平均发病率，频率分布更加趋向于正态分布，且人口基数均达到5 000人以上，发病率分布更为稳定。新构建区域存在一个最大值（100.10/10万）与其他发病率数值相差较大，除掉最大值后，剩余发病率数值中的最大值为51.73/10万，标准差从原来的14.10降低为11.58，该发病率所在的地区可能存在较高的甲状腺癌症的环境风险，应引起相关工作人员的重视和科研工作者进一步的探索和讨论。通过新旧空间尺度的对比分析，发现癌症的发病率在新的尺度上具有较高的稳定性和可靠性，更利于制作癌症发病率分布图及进行相关的空间聚集性分析。对于其他研究区类似的癌症空间聚集性分析具有较大的借鉴和参考意义。

近几十年来，在全世界范围内均观察到甲状腺癌发病率有明显的上升趋势。2005—2009年全国甲状腺癌发病率以每年10.68%的速率增加[14]。引起癌症发病的因素十分复杂，目前唯一明确的会引起甲状腺癌环境风险因素是电离辐射暴露，其他饮食、生活习惯、内分泌干扰物暴露等因素的影响尚无定论[15]，急需对甲状腺癌的时空分布规律进行探测，分析引起其发病率上升的主要原因，进行相应的预

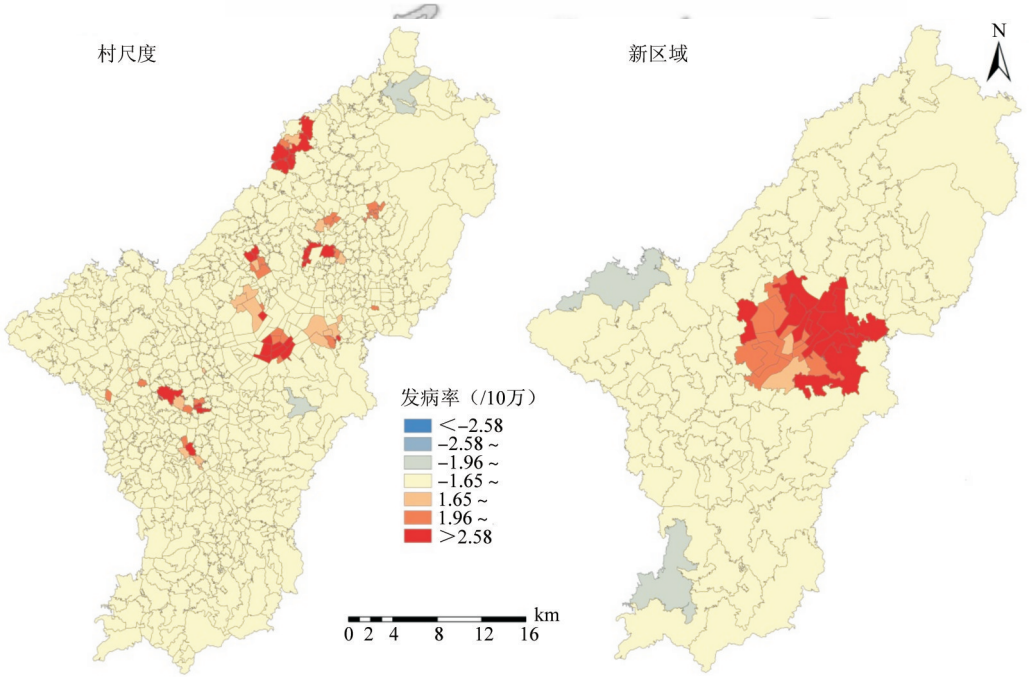


图4　村尺度/新区域甲状腺癌发病率热点探测（Getis-Ord）分布

防和控制。本研究通过产生的新区域，较稳定的发病率分布图，同时通过热点探测发现研究区的中部平原地区有显著的高发病聚集现象。为下一步在高发/低发区进行相应的对照研究和采样分析提供了参考，探索引起甲状腺癌聚集的风险因素奠定了研究基础。

本研究的不足之处在于没有进行甲状腺癌发病率的时间格局分析，由于研究区较小，总体人口基数不足等问题，如果以一年为时间跨度，整个研究区内只能产生29个满足要求的区域（人口＞20 000），大大降低了空间的分辨率，如果在29个区域内进行时空格局分析，会产生精度过低检测不准的现象。为提高研究精度，本研究采用了4年平均的发病率进行分析，势必造成时空格局分析无法进行。但是，本研究介绍的区域化方法适用于其他研究区或者更大范围内的研究，同时，进行了村落的合并，适当降低监测数据发布时的敏感性问题。

参 考 文 献

[1] Guo DS, Wang H. Automatic region building for spatial analysis [J]. T GIS, 2011, 15 Suppl 1: S29-45.

[2] Qi XP, Zhou MG, Hu YS, et al. Spatial hotspot exploration on digestive tract cancer mortality with geographic information system[J]. Geogr Res, 2010, 29(1): 181-187. (in Chinese)
戚晓鹏，周脉耕，胡以松，等. 应用地理信息系统探测消化道癌症死亡率空间聚集性[J]. 地理研究，2010，29(1)：181-187.

[3] Wang FH, Guo DS, McLafferty S. Constructing geographic areas for cancer data analysis: a case study on late-stage breast cancer risk in Illinois[J]. Appl Geogr, 2012, 35(1/2): 1-11.

[4] Lam NSN, Liu K. Use of space-filling curves in generating a national rural sampling frame for HIV/AIDS research [J]. Profess Geogr, 1996, 48(3): 321-332.

[5] Alexander FE, Boyle P. Methods for investigating localized clustering of disease [M]. Lyon, France: IARC Scientific Publications, 1996.

[6] Grady SC, Enander H. Geographic analysis of low birthweight and infant mortality in Michigan using automated zoning methodology[J]. Int J Health Geogr, 2009, 8: 10.

[7] Guo D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP) [J]. Int J Geogr Inf Sci, 2008, 22(7): 801-823.

[8] Su SL. Watershed ecosystem dynamics in response to urbanization: A case of Qiantang River watershed in Zhejiang province [D]. Hangzhou: Zhejiang University, 2013. (in Chinese)
苏世亮. 流域生态系统对城市化的时空响应[D]. 杭州：浙江大学，2013.

[9] Cong MZ, Ou XJ, Zhao Q, et al. Division of land use degree in Jiangsu province based on principal component analysis [J]. Geogr Res, 2008, 27(3): 574-582. (in Chinese)
丛明珠，欧向军，赵清，等. 基于主成分分析法的江苏省土地利用综合分区研究[J]. 地理研究，2008，27(3)：574-582.

[10] Everitt BS, Skrondal A. The Cambridge dictionary of statistics [M]. Cambridge: Cambridge University Press, 2002.

[11] Getis A, Ord JK. The analysis of spatial association by use of distance statistics[J]. Georg Anal, 1992, 24(3): 189-206.

[12] Xiao R, Su SL, Wang JQ, et al. Local spatial modeling of paddy soil landscape patterns in response to urbanization across the urban agglomeration around Hangzhou Bay, China [J]. Appl Geogr, 2013, 39: 158-171.

[13] Ge H, Zhou MG, Wang XF, et al. Applications of multi-level bayes model on Lingbi esophageal cancer mortality spatial distribution pattern research[J]. Chin J Dis Control Prev, 2013, 17(6): 534-537. (in Chinese)
葛辉，周脉耕，王晓风，等. 多水平贝叶斯模型在灵璧县食道癌死亡率空间分布模式研究中应用[J]. 中华疾病控制杂志，2013，17(6)：534-537.

[14] Fei XF, Yang DX, Kong Z, et al. Thyroid cancer incidence in China between 2005 and 2009 [J]. Stoch Environ Res Risk Assess, 2014, 28(5): 1075-1082.

[15] Zhang ZH, Su SL, Xiao R, et al. Identifying determinants of urban growth from a multi-scale perspective: A case study of the urban agglomeration around Hangzhou Bay, China [J]. Appl Geogr, 2013, 45: 193-202.

（收稿日期：2015-03-10）
（本文编辑：王岚）