·流行病学与统计学方法·

空间统计分析方法在结核病研究中的应用

黄茹　李鑫尧　肖洪　田怀玉

【摘要】 传染病的传播与流行是一个复杂的过程，如何定量描述传染病时空分布、分析其与周围环境的关系，是当下研究的重要问题。在结核病的研究中，空间统计分析方法的使用不仅能探索疾病的时空分布模式、构建其与环境变量的统计关系，同时能探测疾病的潜在风险区，为结核病的预防控制提供科学依据。本文将就空间统计分析方法在结核病研究中的应用进行介绍。

【关键词】 结核；　统计学(主题)；　传染病

Application of spatial statistical analysis in tuberculosis study

HUANG Ru, LI Xin-yao, XIAO Hong, TIAN Huai-yu. College of Resources and Environment Science, Hunan Normal University, Changsha 410081, China
Corresponding author: XIAO Hong, Email: xiaohong.hnnu@gmail.com

【Abstract】 The spread and prevalence of infectious diseases is a complicated process. How to quantitatively describe the temporal and spatial distribution of infectious diseases, as well as to analyze the relationship between them and the surrounding is a hotspot for current research. In the study of tuberculosis, spatial statistical analysis could be used to explore patterns of temporal and spatial distribution, to form statistical relationship between them and environmental variables, and predict potential risk factors, all of which are helpful to tuberculosis control and prevention. This study aimed to describe the applications of spatial statistical analysis in tuberculosis research.

【Key words】 Tuberculosis;　Statistics as topic;　Communicable diseases

据估计，约 80%的流行病学研究和公共卫生决策与地理空间信息有关。宿主动物及人群的感染和发病，传播媒介的分布，气温、湿度、降雨、土地利用类型、医疗卫生设施的布局等都具有空间属性[1]。为了更加准确地探索疾病时空分布、制作疾病风险图，以及对疾病与环境因素的相关性进行定量分析，地理信息系统(geographic information system)、遥感(remote sensing)技术及空间统计分析方法被引入该研究领域。这不仅为疾病数据的采集、管理和分析提供了有力工具，也为认识和理解环境变量与发病的数学关系提供了新方法[2-5]。笔者将结合实例，对常用空间统计分析方法及其在结核病研究中的应用进行介绍。

一、空间自相关

1. 方法原理与解释：空间自相关(spatial autocorrelation)是指空间位置上距离越近的事物或现象越相似，通常由空间自相关系数度量[6]。空间正相关是指事物或现象的属性分布具有相似的趋势和取值；若其属性分布具有相反的趋势和取值，则为负相关[7]。空间自相关分析包括全局分析和局域分析，全局空间自相关描述某现象的整体分布状况，区域空间自相关则用来分析局域空间事物或对象的分布是否具有自相关性，局域空间自相关表现出空间聚集性，即空间热点区域[8]。

常用的空间自相关指标有 Moran's I 统计量，Geary's C 比值和 Getis G 统计量等。Moran's I 统计量分为全局 Moran's I 统计和局域 Moran's I 统计，取值范围 −1～1，正值为正相关，负值为负相关，且绝对值越大自相关性越强，0 值表示空间事物分布是随机的，不存在空间自相关[8]。Geary's C 比值范围 0～2，越接近 0 表示空间正相关性越强；越接近 2 则表示空间负相关性越强，越接近 1 表示事物或现象不具有空间相关性[9]。Moran's I 和 Geary's C 统计量均可以用来表明属性值之间的相似程度及在空间上的分布模式，但他们并不能区分是高值的空间集聚[高值簇或热点(hot spots)]，还是低值的空间集聚[低值簇或冷点(cold spots)]，有

doi:10.3969/j.issn.1000-6621.2016.06.003

基金项目：中央高校基本科研业务专项资金(2015NT06)；湖南省科技计划项目(2015JC3063)；湖南省重点学科建设项目(2008001)

作者单位：410081 长沙，湖南师范大学资源与环境科学学院(黄茹、李鑫尧、肖洪)；北京师范大学全球变化与地球系统科学研究院(田怀玉)

通信作者：肖洪，Email：xiaohong.hnnu@gmail.com

可能掩盖不同的空间集聚类型。Getis G 统计量可以识别这两种不同情形的空间集聚[10-11]。计算公式如下：

$$G(d)=\frac{\sum\sum w_{ij}(d)x_i x_j}{\sum\sum x_i x_j} \quad ①$$

x_i 和 x_j 分别表示第 i 个和第 j 个空间位置上的观测值，$w_{ij}(d)$是根据距离规则定义的空间权重。对 Getis G 统计量的统计检验采用下式：

$$Z=\frac{G-E(G)}{\sqrt{Var(G)}} \quad ②$$

检验水准 $\alpha=0.05$。当 $G>0$，且 $P<0.05$ 时，表明观测值之间呈现高值集聚；当 $G<0$，且 $P<0.05$ 时，表明观测值之间呈现低值集聚。

2. 实例介绍：目前常用空间自相关分析来研究疾病时空分布模式。Wang 等[12]使用空间自相关分析，从宏观尺度分析了 2005—2011 年中国大陆地区丙型肝炎病毒(hepatitis C virus，HCV)的地理分布及变换模式，结果表明，HCV 的感染并不是随机分布的，中国中部和边缘(河南、河北、北京、天津和吉林)是感染 HCV 的热点区域。Armién 等[13]利用空间插值和局域空间自相关，结合时间序列啮齿动物监测数据分析啮齿动物的空间行为模式，探测鼠类热点区域，结果显示，鼠类的热点区域很容易随时间发生改变。

3. 结核病研究领域应用介绍：基于不同的时空尺度，空间自相关分析常用于探索结核病时空分布。山珂等[14]对 2002—2011 年全国肺结核疫情进行分析，发现肺结核在全国具有聚集性，肺结核的高高(HH)聚集地区为新疆、贵州、广西地区，低低(LL)聚集地区为北京、天津、河北、山东、辽宁、江苏等地区。康万里[15]通过全局空间自相关分析，发现肺结核发病率和死亡率存在空间聚集性；通过局部自相关分析，发现西藏周边区域是肺结核的高发“热点”区域，上海市周边为低发区域。有学者运用全局 Moran's I 和局部 Moran's I 对浙江省 2000—2011 年肺结核数据进行分析，发现浙江省结核病的 HH 聚集地区位于浙江省南部，LL 地区位于浙江省东北部[16-18]。郑剑等[19]用全局自相关分析及局部自相关分析对湖南省 2012 和 2013 年肺结核患者进行分析，结果显示，肺结核患者分布存在空间聚集性，并用局部自相关分析的方法探索肺结核分布的 HH 聚集区域。Ibrahim 等[20]结合全局自相关分析和局部自相关分析确定尼日利亚结核病发病率的空间格局及聚集性，基于反距离权重的局部自相关分析确定了高低聚集区以及热点的空间位置。

二、时空扫描统计量

1. 方法原理与解释：时空扫描统计量(spatial scan statistic)运用一系列扫描圆探测研究区疾病在时间、空间或时空分布上是否存在聚集性[21]。根据时间、空间维度不同，可分为时间扫描统计量(temporal scan statistic)、空间扫描统计量(spatial scan statistic)和时空扫描统计量(space-time scan statistic)。

2. 实例介绍：时空聚集性分析常用来探索疾病的时空聚集性区域。Zhang 等[22]收集整理了 2005—2012 年中国县级肾综合征出血热(hemorrhagic fever with renal syndrome，HFRS)患者数据，通过局域空间自相关分析和 Kulldorff 时空扫描统计量探索 HFRS 患者的时空动态模式。研究发现，HFRS 的发生具有明显空间正自相关，6、11 和 12 月是 HFRS 的多发季节，中国东北、中部和东部是 HFRS 高发区域。Wu 等[23]结合空间自相关分析和时空聚类分析，探索了辽宁省 HFRS 患者时空聚集性分布，结果发现，HFRS 患者不是随机分布，存在聚集性。

3. 结核病研究领域应用介绍：时空扫描统计量在肺结核空间聚集性分析中已经有较为广泛的应用。康万里和郑素华[24]运用空间扫描统计分析中国菌阳肺结核病患者分布，发现湖南、湖北、江西、四川、广西、广东等省份是菌阳肺结核高发聚集区；低发聚集区主要覆盖北京、天津、山东等地。刘云霞等[25]应用时空重排扫描统计量对青岛市 2006—2007 年结核病患者进行分析，确定 2006—2007 年间青岛市可能存在 5 个结核病聚集区域。裴姣等[26]运用 Turnbull 方法对四川省结核病发病情况进行聚集性分析，结果显示，四川东部、西部和北部是结核病的高发区域。Touray 等[27]用空间扫描统计量对 Greater Banjul 地区肺结核发病数据进行分析，发现发病具有聚集性，并且结核病高发地区的患者大部分是常住居民。Couceiro 等[28]对 2004—2006 年葡萄牙肺结核数据进行聚集性和回归分析，结果表明，一些肺结核的高风险区域的发病率高于 HIV/AIDS 的发病率，贫困率、失业率较高以及总人口较多的地区肺结核的发病风险较高。张英杰等[29]基于地级市空间尺度对全国结核病进行聚集性分析，发现结核病的分布可能具有空间聚集性，黑龙江、吉林和辽宁等省所辖城市结核病流行情况最严重。

三、空间回归模型

空间回归模型在传统统计模型的基础上，考虑

空间相关性、位置和距离[30]，已广泛应用于物种时空分布格局、疾病传播等诸多研究领域[31-34]。

1. 方法原理与解释：地理加权回归模型（geographical weighted regression，GWR）是一种非参数局部线性回归方法，其模型表达式为：

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{k}\beta_j(u_i, v_i)x_{ij} + \varepsilon_i, i = 1, 2, \cdots, n \qquad ③$$

式中(u_i, v_i)为第 i 格中心点坐标；β_j 是随空间地理位置变化的回归系数；ε_i 为独立同分布的误差项。该模型的回归系数是区域地理位置的函数，并随地理位置的变化而变化，并用以探索空间数据的空间异质性，因此其回归结果更加可信[35]。

规则集遗传算法（genetic alorithm for rule-set production，GARP）模型则是一种求解最优参数组合的 GWR 模型。GARP 模型利用患者点位数据和环境集数据，通过反复迭代形成由不同规则共同组成的模型，用以表示物种的生态需求，探索物种分布和研究区环境因子之间的非随机关系[36]。在建模过程中，患者或宿主动物分布数据被随机均分为训练数据和测试数据，测试数据用于外部检测评价，不参与模型的构建[37]。但由于 GARP 模型不稳定，要选择一定数量的候选模型生成物种分布的等级图[38]。

2. 实例介绍：传统回归模型在探测疾病空间风险中仍然被大量使用。Si 等[39]应用 logistic 回归模型分析了欧洲家禽中暴发 H5N1 与环境因素的关系，其中温度、降雨和湿地地区的人口密度等是影响 H5N1 高致病性禽流感传播的主要因素。Xiao 等[40]采用 logistic 模型，结合归一化植被指数（normalized difference vegetation index，NDVI）、温度植被干旱指数（temperature vegetation drought index，TVDI）以及相关环境变量分析湖南省四市两县（长沙、衡阳、湘潭、株洲、双峰县和邵东县）、汉坦病毒感染的生态环境特征，结果表明，汉坦病毒感染风险主要发生在 TVDI 较大而海拔较低的区域。肖洪等[41]结合时空聚类分析与泊松回归分析，探索 HFRS 传播的时空分布与地理景观影响因素，结果显示，HFRS 患者呈时空聚集性分布；HFRS 发病风险随着耕地面积的增大而增加，随着林地、农村居民点面积的增加而降低。

近年来，在针对疾病和宿主动物的地理分布格局的探索研究中，GARP 模型得到了广泛应用。Xiao 等[42]利用 GARP 等生态位模型获取了湖南省长沙市 HFRS 在不同区域传播的生态环境特征，发现 HFRS 发病风险集中在海拔低于 200 m、年平均气温 17.5 ℃、年降水量不足 1600 mm 和 NDVI 较低的区域。在针对湘江中下游地区 HFRS 发病风险区的研究中，NDVI 和土地利用对 HFRS 传播有重要影响，城镇和建筑用地是 HFRS 的主要风险用地类型[34]。Haredasht 等[43]应用 GARP 生态位模型对西欧银行田鼠的生态特征进行了研究，结果发现，田鼠在最热的季节分布在降水为 300～550 mm 的区域，而在最冷的季节分布在温度为－5～－10 ℃区域。

然而，应用生态位模型仍有很多亟待解决的问题：（1）人类活动是影响生物分布的重要因素，在生态位模型研究中如何综合考察人类活动对物种生态位的影响；（2）如何根据具体问题选择复杂度适宜的模型结构；（3）理解环境变量的物理意义、生物学假设，以及算法参数的设置和模型的评价是未来研究的重点。

3. 结核病研究领域应用介绍：吴田勇等[44]采用空间误差模型对重庆市结核病空间分布的影响因素进行讨论，结果表明，结核病发病只与城镇失业率呈正相关。Munch 等[45]通过泊松回归和 K-均值聚类对 1993—1996 年 Ravensmead 和 Uitsig 地区的肺结核高发区域进行分析，结果显示，失业率和经济贫困程度与结核病的高发病率呈正相关。刘云霞等[46]构建 GWR 模型探索山东省结核病及其影响因素间的局域关系，结果表明，不同区域各影响因素对结核病登记率的影响存在程度和方向上差异。

综上，尽管空间统计分析方法已广泛应用于结核病时空分布模式研究，但少有研究全面探索了结核病发生与环境因素的空间统计关系。由于结核病数据处理过程中存在误差及数据共享性差，加之空间统计学（模型、技术和方法）的学科特点及研究者缺乏专业知识，这些都制约了计量地理学在结核病数据处理和分析中应用。结核病的发生和发展是一个复杂的过程，对其时空分布格局进行准确可靠的分析，需要不断积累新知识，探索新方法，以不断提高定量分析和预测的准确性，为结核病的理论研究及预防控制提供有力的工具和技术支持。

参 考 文 献

[1] Hallett TB, Coulson T, Pilkington JG, et al. Why large-scale climate indices seem to predict ecological processes better than local weather. Nature, 2004, 430(6995): 71-75.
[2] 肖洪，田怀玉. 传染病时空传播研究：规律探索与决策支持. 中华预防医学杂志，2012，46(6)：492-494.
[3] 杨瑞馥. 传染病防控研究：机遇与挑战. 中华预防医学杂志，

2011,45(10): 869-872.
[4] 赖圣杰，廖一兰，张洪龙，等. 2011—2013 年国家传染病自动预警系统中时间模型和时空模型应用效果比较. 中华预防医学杂志，2014,48(4): 259-264.
[5] 李中杰，马家奇，赖圣杰，等. 2011—2013 年国家传染病自动预警系统运行结果分析. 中华预防医学杂志，2014,48(4): 252-258.
[6] Cliff AD, Ord JK. Spatial autocorrelation. London: Pion, 1973.
[7] 孙果梅. 流动人口对上海市结核病疫情及发病模式的影响. 上海：复旦大学，2012.
[8] 王劲峰，廖一兰，刘鑫. 空间数据分析教程. 北京：科学出版社，2010.
[9] 冯军，吴晓华，李石柱，等. 空间统计分析方法及相关软件在传染病研究中的应用. 中国血吸虫病防治杂志，2011,23(2): 217-220.
[10] Getis A, Ord J. The analysis of spatial association by use of distance statistics. Geogr Anal, 1992,24(3): 189-206.
[11] Ord J, Getis A. Local spatial autocorrelation statistics: distributional issues and an application. Geogr Anal, 1995,27(4): 286-306.
[12] Wang L, Xing J, Chen F, et al. Spatial analysis on hepatitis C virus infection in mainland China: from 2005 to 2011. PLoS One, 2014,9(10): e110861.
[13] Armién B, Ortiz PL, Gonzalez P, et al. Spatial-temporal distribution of hantavirus rodent-borne infection by *oligoryzomys fulvescens* in the Agua Buena Region-Panama. PLoS Negl Trop Dis, 2015,10(2): e0004460.
[14] 山珂，徐凌忠，盖若琰，等. 中国 2002—2011 年肺结核流行状况 GIS 空间分析. 中国公共卫生，2014,30(4): 388-391.
[15] 康万里. 空间分析方法在中国结核病分布和 120 急救系统中的应用. 太原：山西医科大学，2007.
[16] 柴鹏飞. 鄞州区 2005—2008 年肺结核病疫情的空间统计分析. 杭州：浙江大学，2009.
[17] 桂娟娟，张添方，刘志芳，等. 浙江省 2005—2011 年肺结核流行特征与空间聚集性. 中国公共卫生，2016,32(1): 11-14.
[18] 颜梦欢，丁海峰，马海燕. 浙江省 2000—2011 年肺结核流行状况及空间自相关分析. 中国公共卫生，2015,31(1): 25-28.
[19] 郑剑，唐益，查文婷，等. 湖南省 2012 和 2013 年肺结核 GIS 空间流行病学分析. 中国公共卫生，2015,31(12): 1590-1593.
[20] Ibrahim S, Hamisu I, Lawal U. Spatial pattern of tuberculosis prevalence in nigeria: a comparative analysis of spatial autocorrelation indices. American Journal of Geographic Information System, 2015,4(3): 87-94.
[21] Kulldorff M. A spatial scan statistic. Commun Stat-Theor M, 1997,26(6): 1481-1496.
[22] Zhang WY, Wang LY, Liu YX, et al. Spatiotemporal transmission dynamics of hemorrhagic fever with renal syndrome in China, 2005—2012. PLoS Negl Trop Dis, 2014, 8(11): e3344.
[23] Wu W, Guo J, Guan P, et al. Clusters of spatial, temporal, and space-time distribution of hemorrhagic fever with renal syndrome in Liaoning Province, Northeastern China. BMC Infect Dis, 2011,11:229.
[24] 康万里，郑素华. 空间扫描统计在中国菌阳结核病分布中的应用. 中国卫生统计，2012,29(4): 298-300.
[25] 刘云霞，李士雪，王忠东，等. 基于时空重排扫描统计量的结核病聚集性研究. 山东大学学报：医学版，2009,47(12): 122-125.
[26] 裴姣，殷菲，李晓松，等. Turnbull 方法在四川省结核病空间聚集性分析中的应用初探. 中华疾病控制杂志，2011,15(5): 441-444.
[27] Touray K, Adetifa IM, Jallow A, et al. Spatial analysis of tuberculosis in an urban west African setting: is there evidence of clustering? Trop Med Int Health, 2010,15(6): 664-672.
[28] Couceiro L, Santana P, Nunes C. Pulmonary tuberculosis and risk factors in Portugal: a spatial analysis. Int J Tuberc Lung D, 2011,15(11): 1445-1454.
[29] 张英杰，曹凯，王超，等. 中国结核病地级市水平空间积聚性分析. 现代预防医学，2015,42(17): 3089-3092.
[30] Anselin L. Spatial econometrics: methods and models. Dordrecht: Kluwer Academic, 1988.
[31] Monahan WB, Tingley MW. Niche tracking and rapid establishment of distributional equilibrium in the house sparrow show potential responsiveness of species to climate change. PLoS One, 2012,7(7): e42097.
[32] Wang TL, Wang GY, Innes J, et al. Climatic niche models and their consensus projections for future climates for four major forest tree species in the Asia-Pacific region. Forest Ecol Manag, 2016,360: 357-366.
[33] 唐启强，张智，赵安，等. 鄱阳湖区南昌县血吸虫疫情空间分布及其流行因素分析. 热带地理，2013,133(1): 76-80.
[34] 肖洪，林晓玲，高立冬，等. 湘江中下游肾综合征出血热传播风险预测和环境危险因素分析. 地理科学，2013,33(1): 123-128.
[35] 黄秋兰，唐咸艳，周红霞，等. 应用空间回归技术从全局和局部两个水平上定量探讨影响广西流行性乙型脑炎发病的气象因素. 中华疾病控制杂志，2013,17(4): 282-286.
[36] Stockwell D, Peters D. The GARP modeling system: problems and solutions to automated spatial prediction. Int J Geogr Inf Sci, 1999,13(2): 143-158.
[37] 林晓玲，肖洪，田怀玉. 生态位模型在传染病风险预测中的应用. 中华预防医学杂志，2013,47(4): 294-296.
[38] 李双成，高江波. 基于 GARP 模型的紫茎泽兰空间分布预测——以云南纵向岭谷为例. 生态学杂志，2008,27(9): 1531-1536.
[39] Si YL, de Boer WF, Gong P. Different environmental drivers of highly pathogenic avian influenza H5N1 outbreaks in poultry and wild birds. PLoS One, 2013,8(1): e53362.
[40] Xiao H, Huang R, Gao LD, et al. Effects of humidity variation on the hantavirus infection and hemorrhagic fever with renal syndrome occurrence in subtropical China. Am J Trop Med Hyg, 2016,94(2): 420-427.
[41] 肖洪，田怀玉，代翔宇，等. 地理景观对长沙市肾综合征出血热传播的影响. 中华预防医学杂志，2012,46(3): 246-251.
[42] Xiao H, Lin X, Gao L, et al. Ecology and geography of hemorrhagic fever with renal syndrome in Changsha, China. BMC Infect Dis, 2013,13:305.
[43] Amirpour Haredasht S, Barrios M, Farifteh J, et al. Ecological niche modelling of bank voles in Western Europe. Int J Environ Res Public Health, 2013,10(2): 499-514.
[44] 吴田勇，曾庆，刘世炜，等. 重庆市 2008—2011 年结核病疾病空间分布及影响因素分析. 上海交通大学学报：医学版，2013,33(4): 489-492.
[45] Munch Z, Van Lill SW, Booysen CN, et al. Tuberculosis transmission patterns in a high-incidence area: a spatial analysis. Int J Tuberc Lung Dis, 2003,7(3): 271-277.
[46] 刘云霞，刘言训，张冰冰，等. 基于 GWR 模型的结核病空间流行病学研究. 中国防痨杂志，2013,35(5): 343-346.

（收稿日期：2016-04-14）
（本文编辑：李敬文）