学习·发现·交流

时空统计方法

樊文洁 王山 曹红艳 江涛 李秀央

【关键词】 时空统计；空间测距法；集结法；空间插值分析；聚类检测技术；多元空间回归分析

Spatio-temporal statistical method *Fan Wenjie*[1], *Wang Shan*[1], *Cao Hongyan*[2], *Jiang Tao*[3], *Li Xiuyang*[1]. 1 *Department of Epidemiology and Biostatistics, Medical College, Zhejiang University, Hangzhou* 310058, China; 2 *Jianggan District Center of Disease Control and Prevention, Zhejiang*; 3 *Zhejiang Provincial Center of Disease Control and Prevention*

Corresponding author: *Li Xiuyang*, Email: lixiuyang@zju.edu.cn

This work was supported by grants from the University Scientific Research Plan in 2012 of Department of Education of Zhejiang Province (No. Y201225233) *and Science and Technology Innovation and Planted Talent Plan Projects for College Students in Zhejiang Province in 2014* (No. 2014R401101).

【Key words】 Spatio-temporal statistics; Spatial proximity; Aggregation methods; Spatial interpolation; Cluster detection techniques; Multivariable spatial regression analysis

近年来，应用空间统计分析方法进行疾病分布图描述、疾病聚集性研究、地理环境与疾病相关性研究，以及疾病危险因素研究逐渐成为热点。由于空间分析方法的日益丰富和局部地理数据可获渠道的多样性，现代空间流行病学已成为流行病学分支，特别适用于局部区域的空间分析，主要应用于环境污染监测、传染病和慢性疾病的预防控制、公共卫生应急管理和预警技术3个方面[1]。本文复习空间统计分析中主要几种方法。

1. 空间测距法（spatial proximity）：该法的基本思想是通过测算目标地点（如住宅和工作场所）到目标资源或危害源的距离以解释暴露的大小。一般用于人群行为模式或环境污染的研究，如研究空气污染一般通过计算目标地点到最近的铁路距离以评估环境污染状况[2]；行为学家通过测算研究地点到最便捷的快餐店或零食商店，研究人群饮食结构的差异[3]，测算到公园绿地等的距离以研究人群运动习惯的差异[4]，测算到药店的距离以研究患病求医行为的差异[5]。

在此所用的距离一般为点之间的直线距离，尽管很方便，但往往非常粗糙，不能反映真实情况，故有学者建议使用网状路线距离来计算实际距离以减小误差[5]。当然，若两点之间距离较远而致出行时间成为掣肘，测距时也应考虑其出行方式。测距法的最大优势在于其原理简单，应用方便，劣势在于误差过大。

2. 集结法（aggregation methods）：常用以评估暴露的一种空间方法。其基本原理是用给定空间（如一个普查区）的某一特征量的合计量或平均水平来评估某一特征，如以某地区车辆的总数评估该地区的交通污染暴露情况[5-6]，或以某城市的道路交叉路口数评估该城市的环境宜步行性[7]。该法常与空间测距法联用，通过加入距离权重进行资源评估。使用中可能遇到困难，如研究特定空间中基础人口分布对变异的影响，人口密度会影响到人均可用资源，即更高的人口密度意味着对资源愈加激烈的竞争性。目前有两种方法可以解决该问题：第一，研究空间的大小可以根据人口密度进行调整；第二，研究人员可以使用一个固定的缓冲区，其密度由人口数量加权[8-9]。虽然缓冲区域集结法应用简便，但仍太复杂，即当缓冲区域作为部分测量区域时，需要使用修正因子；由于相关的缓冲区大小未知，需要对可替代的缓冲区大小作灵敏度分析[10]。

集结法也常用ArcGIS进行，考虑到人口密度的影响时，使用全球人口动态统计数据库LandScan。该数据库由美国能源部橡树岭国家实验室（ORNL）开发，East View Cartographic提供。LandScan运用GIS和遥感等创新方法，是全球人口数据发布的社会标准，也是全球最为准确、可靠，具有分布模型和最佳分辨率的全球人口动态统计分析数据库。

3. 空间插值分析（spatial interpolation）：空间流行病学也常用抽样研究。利用样本点值的空间分布规律可以对未抽样点值进行估计，估计值可以制作疾病地图，供卫生决策参考。由于空间插值分析是通过有限的样本点数据，对地图平面上所有点位值进行估计，采用这些估计值所制作的疾病地图可以连成一个光滑的表面[11]，故空间插值分析又被认为是一种平滑技术。该法主要有距离倒数插值、趋势面分析、样条插值、权重插值、Kriging插值等方法。前两种方法现在应用较少，在研究中一般使用后三种。

（1）距离倒数插值：该法是基于两空间位置属性的相似性或相关性与距离成反比，即距离越远，影响越小。

DOI：10.3760/cma.j.issn.0254-6450.2015.01.019

基金项目：浙江省教育厅2012年度高校科技计划（Y201225233）；2014年浙江省大学生科技创新活动暨新苗人才计划（2014R401101）

作者单位：310058 杭州，浙江大学医学院流行病与卫生统计学系（樊文洁、王山、李秀央）；杭州市江干区疾病预防控制中心（曹红艳）；浙江省疾病预防控制中心（江涛）

通信作者：李秀央，Email：lixiuyang@zju.edu.cn

（2）趋势面分析：是用以研究区域尺度上空间结构的趋势和逐渐变化的一种空间分析方法。其基本思想是将数据的空间变化分解成3个部分：区域趋势、局部异常和随机误差，其实质是进行数据拟合。趋势面本身是一个多项式函数，随多项式次数提高，虽然拟合程度越高，但其能用性和预测性也就越低，计算也越复杂，实际应用中通常采用的是二阶或三阶多项式函数。

（3）样条插值：是使用一种称为样条的特殊分段，采用多项式进行插值，以解决低阶插值函数拟合程度差，高阶插值函数计算量大、有剧烈振荡、数值稳定性差，分段线性插值在分段点上仅连续而不光滑等问题。该方法特别适用于局部地区趋势明显的数据研究，并不要求数据分布，计算量也较小[12-13]。

（4）权重插值：权重插值可综合考量空间面积大小、形状影响、人口数量、信息变异速度等各影响量，其优点是计算量小。

（5）Kriging插值：该方法由Krige于1951年提出。其原理是空间距离相关和方向相关，在数学上被证明是空间分布数据局部最优线性无偏估计技术[14]。线性是指估计值是样本值的线性组合；无偏是指估计值的数学期望等于理论值，最优是指估计的误差方差最小。一些研究中常用地理系统基础的Kriging插值估计患病率[15-16]。

4. 聚类检测技术（cluster detection techniques）：在非随机分布的空间数据中，空间聚类分析是最常用的工具。其中有些用以检验一些疾病地理空间聚类及其聚集是否具有偶然性。根据研究范围的大小和明确的研究地区，又分为全球聚类检验、局部聚类检验和焦点聚类检验3种类型。

全球集类检测可确定未给限定地点的大区域存在的聚类，最常用的方法有Diggle Chetwynd二元K检验、Mantel-Bailar测试、Potthoff-Whittinghill（PW）测试，较少用的方法有Moran’s I统计，至近邻近法（Cuzick and Edwards’s nearest neighbors）和最大化额外事件测试（Tango’s maximized excess events tests，MEET），其中以MEET的检验效能最强，可评估空间自相关和空间异质性[17]。局部聚类检验也称热点分析，用于检验特定小规模区域的聚集性。常用方法有空间相关性的局部指标法（Anselin’s local indicator of spatial association，LISA）、Besag-Newell检验法、时空扫描统计量。Naus提出了扫描统计量的概念，即用事先选定的时间区间扫描整个观察期得到的病例数最大值。由于该方法消除了人为地按年或月分组造成的主观性，而且检验效能较高，可以调整为多联性测试，允许非均匀人口密度的背景，及其他混杂变量，适用于点数据和聚集性数据，与SaTScan相结合，目前已成为疾病时间聚集性或区域聚集性分析的热点[18-19]。根据资料性质不同，常用的主要方法有Bernoulli模型的扫描统计量、Poisson模型的扫描统计量、时空重排模型的扫描统计量、Ordinal模型的扫描统计量和指数模型的扫描统计量等[20]。

尽管大多数疾病调查的集群本质上是空间的，而研究探索癌症和传染病往往涉及到时空分析，许多研究应用Knox全球时空分析技术和Diggle全球时空K检验技术[21-22]，其中K检验更佳，因为即可更好纠正边缘效果也可允许更大范围的时空尺度。其他方法还包括研究当地时空聚类的空间扫描统计量、用于前瞻性识别和监测高危地区的连续时空系统（dynamic continuous-area space-time system，DYCAST）、用于数据不确定性的广义Bayes最大熵技术（generalized Bayesian maximum entropy，GBME）和用于探索流行波在长期时段速度的空间速度技术（spatial velocity techniques）等。

5. 多元空间回归分析（multivariable spatial regression）：空间回归模型是生态学分析的主要方法，从生态学的角度研究疾病发病（或患病、死亡等）空间分布与解释变量（环境因素，如空气、水、土壤等及社会经济学因素）间的关系。在传统分析中，分析结果变量和解释变量的关系时，常采用线性回归或logistic回归等方法，均要求个体间彼此独立，而由于受共同环境影响，在空间分布的个体间可能彼此相关，故在传统的回归分析中引入随机效应项，以解释可能存在的空间相关性的影响。

标准的统计回归模型，其数据要求具有独立性，因此不适合分析空间数据的分析。空间建模需要对准则和调整后的数据就空间自相关性的强度进行反复评估。如果空间数据具有自相关性或者协变量信息不能完全解释该模式，那么就必须将空间相关性引入模型。尽管有大量的空间和非空间统计模型，但其间的差异很小[12,23]，实际应用中常采用空间自相关分析和Bayes统计模型。

（1）空间自相关分析：空间自相关性是指空间位置上越靠近的事物或现象就越相似，即事物或现象具有对空间位置的依赖关系。空间自相关分析包括全程空间自相关分析和局部空间自相关分析，需要的空间数据类型是点或面数据，分析的对象是具有点或面分布特征的特定属性。空间流行病学中，表示空间自相关大小的常用统计量有3个，即Moran’s I、Geary’s C和G统计量。Moran’s I统计量是一个应用最广的衡量空间自相关性指标，可用来进行全程或局部空间自相关分析；Geary’s C统计量是另一个常用于分析全局空间自相关性的指标；G统计量用来分析局部空间自相关性。

（2）Bayes统计模型：近年来生态学研究的重点已转向针对小区域的空间研究上。但由于区域范围小，计数总量少且分散，变异大，不同区域往往存在空间相关性，Bayes分析能克服这些困难，因而成为疾病分析的主流方法。其基本原理为通过构建分层Bayes模型对未知参数只提出先验分布，并进行Bayes估计获得Bayes后验分布，再通过马尔科夫链-蒙特卡罗方法（Markov Chain Monte Carlo method，MCMC）进行后验分布的计算，最终获得参数的估计值，可以有效对包括小区域范围疾病分布图的描绘、疾病地理聚集性分析和疾病地理相关性研究在内的时空非独立数据进行分析[24]。目前Bayes空间分析模型已发展成多类的分析方法，包括近年来充分发展和得到应用的BYM（Besag York and Mollié）模

型、联合随机模型、半参数Bayes统计及移动性均化模型等。其中，以BYM模型和半参数模型中的MIX模型的优势较明显。

BYM模型是目前最成熟且应用最广泛的方法。在利用固定协变量和随机效应的多数研究中，利用BYM模型可定量单一疾病的潜在风险因子效应[23,25-26]；在部分研究中，利用带有共享组件的BYM模型可建立二元或多元疾病的联合空间模型以分析共同的风险因子[27-29]。另外，还发展了空间BYM模型进行时空分析[27]。在半参数模型中，相对BYM模型而言，MIX模型最受关注，且更加反映真实情况，更具灵活性和适应性。MIX模型倾向于把各个地区的发病资料划分类别，形成危险性地区或者普通地区，因此，MIX模型可以显著发现高危险性地区。而BYM模型往往会中和高危地区和低危地区。但是对于整个区域发病缓和的情况，MIX模型的处理能力却不如BYM模型[20]。

6. 基于空间动态面板模型：这是处理截面数据中空间效应（spatial effects）专门的计量模型和统计方法。该法能同时考虑被解释变量在时间动态（dynamic）效应与在空间上的溢出（spillover）效应。现代的许多领域，例如区域科学[30]、环境经济学[31]、农业经济学、公共财政[32]和流行病学等研究已越来越多地考虑截面数据上的空间自相关性。

7. 统计软件：时空统计分析常用软件见表1。

表1 时空统计分析常用软件

统计软件	应用范围
ArcGIS	地图可视化，空间聚类检测，自相关分析
SaTScan	空间扫描法
SAS	集结法、空间插值技术、聚类检测技术和标准回归
WinBUGS	Bayes建模
GeoDa	自相关性统计和异常值指示
R	提供各种数学计算、统计计算的函数，创造新统计算法

综上所述，空间测距法和集结法由于其方法简便，是空间分析中最为青睐的方法；空间插值分析、聚类检测技术、多元空间回归分析相对应用较少，其中空间扫描统计法和Bayes回归模型法由于其独特的优势，逐渐成为时空分析的热点。在实际应用中，由于空间测距法和集结法偏倚过大，分析问题流于表面和简单，同样也不适用于大规模大数据的研究，近来其应用已趋减少。空间扫面统计法和建立回归模型是建立在大样本数据之上，加之SaTScan等软件的开发方便了运算步骤，其准确性得到可靠保证，越受青睐。然而在流行病学中应用空间分析方法的研究仍为少数，大部分流行病学专家不擅长空间分析方法，但随着计算机技术的飞速发展和各类空间技术水平的不断提高，这一交叉学科必定会推动公共卫生学、生物学（遗传学）、生态学、数学、数理统计学、计算机科学、GIS等专业的发展，并将在医学与公共卫生领域中发挥更大作用。

参 考 文 献

[1] Zhou XN, Yang GJ, Yang K, et al. Progress and trends of spatial epidemiology in China [J]. Chin J Epidemiol, 2011, 32(9): 854-858. (in Chinese)
周晓农，杨国静，杨坤，等. 中国空间流行病学的发展历程与发展趋势[J]. 中华流行病学杂志，2011，32(9)：854-858.

[2] Allen RW, Criqui MH, Diez Roux AV, et al. Fine particulate matter air pollution, proximity to traffic, and aortic atherosclerosis [J]. Epidemiology, 2009, 20: 254-264.

[3] Davis B, Carpenter C. Proximity of fast-food restaurants to schools and adolescent obesity [J]. Am J Public Health, 2009, 99: 505-510.

[4] Zenk SN, Schulz AJ, Israel BA, et al. Neighborhood racial composition, neighborhood poverty, and the spatial accessibility of supermarkets in metropolitan Detroit [J]. Am J Public Health, 2005, 95: 660-667.

[5] Michael YL, Perdue LA, Orwoll ES, et al. Physical activity resources and changes in walking in a cohort of older men [J]. Am J Public Health, 2010, 100: 654-660.

[6] Bailey TC, Gatrell AC. Interactive Spatial Data Analysis [M]. Hoboken, NJ: Wiley, 1995.

[7] Cervero R, Duncan M. Walking, bicycling, and urban landscapes: evidence from the San Francisco Bay Area [J]. Am J Public Health, 2003, 93: 1478-1483.

[8] Berke EM, Tanski SE, Demidenko E, et al. Alcohol retail density and demographic predictors of health disparities: a geographic analysis [J]. Am J Public Health, 2010, 100: 1967-1971.

[9] Chaix B, Merlo J, Subramanian SV, et al. Comparison of a spatial perspective with the multilevel analytical approach in neighborhood studies: the case of mental and behavioral disorders due to psychoactive substance use in Malmo, Sweden, 2001 [J]. Am J Epidemiol, 2005, 162: 171-182.

[10] Diez Roux AV, Evenson KR, McGinn AP, et al. Availability of recreational resources and physical activity in adults [J]. Am J Public Health, 2007, 97: 493-499.

[11] Moore DA, Carpenter TE. Spatial analytical methods and geographic information systems: use in health research and epidemiology [J]. Epidemiol Rev, 1999, 21: 143-161.

[12] Auchincloss AH, Diez Roux AV, Brown DG, et al. Filling the gaps: spatial interpolation of residential survey data in the estimation of neighborhood characteristics [J]. Epidemiology, 2007, 18: 469-478.

[13] Ugarte MD, Goicoa T, Etxeberria J, et al. Age-specific spatio-temporal patterns of female breast cancer mortality in Spain (1975-2005) [J]. Ann Epidemiol, 2010, 20: 906-916.

[14] Goovaerts P. Geostatistics for Natural Resources Evaluation [M]. Oxford: Oxford University Press, 1997.

[15] Gemperli A, Vounatsou P, Kleinschmidt I, et al. Spatial patterns of infant mortality in Mali: the effect of malaria endemicity [J]. Am J Epidemiol, 2004, 159: 64-72.

[16] Gesink Law DC, Bernstein KT, Serre ML, et al. Modeling a syphilis outbreak through space and time using the Bayesian maximum entropy approach [J]. Ann Epidemiol, 2006, 16: 797-804.

[17] Thomas W, Birgit R, Edith S. Changing geographical distribution

of diabetes mellitus type 1 incidence in Austrian children 1989-2005[J]. Eur J Epidemiol, 2008, 23: 213-218.

[18] Kulldorff M. A spatial scan statistic[J]. Commun Stat Part A Theory Methods, 1997, 26: 1481-1496.

[19] Gregorio DI, Huang L, DeChello LM, et al. Place of residence effect on likelihood of surviving prostate cancer[J]. Ann Epidemiol, 2007, 17: 520-524.

[20] Li XY, Chen K. Scan statistics and its application in spatial epidemiology[J]. Chin J Epidemiol, 2008, 29(8): 828-831. (in Chinese)
李秀央, 陈坤. 扫描统计量的理论及其在空间流行病学中的应用[J]. 中华流行病学杂志, 2008, 29(8): 828-831.

[21] Gustafsson B, Carstensen J. Space-time clustering of childhood lymphatic leukaemias and non-Hodgkin's lymphomas in Sweden [J]. Eur J Epidemiol, 2000, 16: 1111-1116.

[22] Houben MP, Coebergh JW, Birch JM, et al. Space-time clustering of glioma cannot be attributed to specific histological subgroups [J]. Eur J Epidemiol, 2006, 21: 197-201.

[23] Boyd HA, Flanders WD, Addiss DG, et al. Residual spatial correlation between geographically referenced observations: a Bayesian hierarchical modeling approach[J]. Epidemiology, 2005, 16: 532-541.

[24] Zheng WJ, Li XY, Chen K. Bayesian statistics in spatial epidemiology[J]. J Zhejiang Univ: Med Sci, 2008, 37(6): 644-649. (in Chinese)
郑卫军, 李秀央, 陈坤. 空间流行病学研究中的Bayes统计方法[J]. 浙江大学学报: 医学版, 2008, 37(6): 644-649.

[25] Dominguez-Berjon MF, Gandarillas A, Segura del Pozo J, et al. Census tract socioeconomic and physical environment and cardiovascular mortality in the region of Madrid (Spain)[J]. J Epidemiol Commun Health, 2010, 64: 1086-1093.

[26] Hickson DA, Waller LA, Gebreab SY, et al. Geographic representation of the Jackson Heart Study cohort to the African-American population in Jackson, Mississippi[J]. Am J Epidemiol, 2011, 173: 110-117.

[27] Feltbower RG, Manda SO, Gilthorpe MS, et al. Detecting small-area similarities in the epidemiology of childhood acute lymphoblastic leukemia and diabetes mellitus, type 1: a Bayesian approach[J]. Am J Epidemiol, 2005, 161: 1168-1180.

[28] Havulinna AS, Paakkonen R, Karvonen M, et al. Geographic patterns of incidence of ischemic stroke and acute myocardial infarction in Finland during 1991-2003[J]. Ann Epidemiol, 2008, 18: 206-213.

[29] Manda SO, Feltbower RG, Gilthorpe MS. Investigating spatio-temporal similarities in the epidemiology of childhood leukaemia and diabetes[J]. Eur J Epidemiol, 2009, 24: 743-752.

[30] Li R, Wang LP, Liu QQ. The empirical analysis on the influence of geographical distance and economic distance to entrepreneurship knowledge spillover[J]. Sci Technol Progress Policy, 2012, 29(10): 113-118. (in Chinese)
李燃, 王立平, 刘琴琴. 地理距离与经济距离对创业知识溢出影响的实证分析[J]. 科技进步与对策, 2012, 29(10): 113-118.

[31] Wei LL, Sun X, Wang LP. Spatial structure dynamic analysis of industrial transfer and environmental pollution[J]. Inquiry into Economic Issues, 2010, 10: 23-27. (in Chinese)
未良莉, 孙欣, 王立平. 产业转移与环境污染的空间动态面板分析[J]. 经济问题探索, 2010, 10: 23-27.

[32] Gong LZ. Influencing factors and regional disparity of growth of China's land remise income — an empirical study based on dynamic model of spatial panel data[J]. Res Economics Management, 2011, 9: 15-23. (in Chinese)
龚丽贞. 我国土地出让收入增长的影响因素与地区差异——基于空间动态面板模型的实证研究[J]. 经济与管理研究, 2011, 9: 15-23.

(收稿日期: 2014-07-13)
(本文编辑: 张林东)

中华流行病学杂志第七届编辑委员会通讯编委名单

(按姓氏汉语拼音排序)