

空间回归模型在公共卫生中的应用

伍劲屹 周艺彪 姜庆五

【关键词】 空间回归模型; 公共卫生; 空间自相关

Application of spatial regression model in public health
WU Jin-yi, ZHOU Yi-biao, JIANG Qing-wu. Department of Epidemiology, School of Public Health, Fudan University, Key Laboratory on Public Health Safety, Ministry of Education, Research Center on Tropical Disease, Fudan University, Shanghai 200032, China

Corresponding author: ZHOU Yi-biao, Email: ybzhou@fudan.edu.cn

This work was supported by grants from the National Natural Science Foundation of China (No. 30872159) and the National Science and Technology Mega-Projects of China (No. 2012ZX10004-220).

【Key words】 Spatial regression model; Public health; Spatial autocorrelation

随着地理信息系统(GIS)、全球定位系统(GPS)、遥感系统(RS)即3S体系的发展,空间数据呈爆炸式增长,而科研人员也更容易从网络上获得信息^[1]。据统计80%的流行病学资料具有空间属性^[2],且有空间自相关性质,采用传统的统计学分析方法并不符合其对数据独立性的要求,但空间回归模型不要求数据独立,可以利用数据的空间属性。在用于传染病分析时,可考虑疾病的聚集性和扩散性,并可说明各因素间的关系。空间回归模型已成功应用于探讨基本时空格局的影响因素以及空间变异性^[3]。为此本文复习空间回归模型的概念和基本原理及其在公共卫生中的应用,并初步探讨如何正确使用空间回归模型。

1. 空间回归模型:

(1)空间回归模型概念和基本原理:1988年出版的*Spatial Econometrics: Methods and Models*一书^[4]提出将空间效应融合到普通线性回归模型中,形成空间回归模型,并对空间回归的一般模型、参数估计和假设检验进行拓展研究。其基本原理是将空间中邻近的资料通常比相离较远的资料具有较高相似性,空间回归分析通过位置建立资料间的统计关系^[5]。

(2)分析过程及其优点:①首先进行单变量分析,计算出最小值、最大值、均数、标准差;②在满足应变量服从正态分布的情况下,用普通最小二乘法(ordinary least square)^[6]进行模型拟合,从而得到残差的Moran'I统计量^[7],判断残差是否有空间自相关性;③做出局部空间自相关性指示变量(local

indicator of spatial autocorrelation)空间聚集性地图^[8]和Moran'I散点图^[9];④存在空间自相关性时,使用各空间回归模型分别进行拟合;⑤通过多个指标拉格朗日指数(Robust Lagrange multiplier index)^[10]、赤池信息量准则(Akaike information criterion)^[11,12]、对数似然值(log likelihood)^[13]等进行模型的选择;⑥通过模型的各系数,对自变量与应变量的关系进行解释。

传统统计模型通常要求变量具有独立性,不考虑空间相关性、位置和距离,这使得其对传染病等具有空间自相关性问题的研究有很大局限性。公共卫生中的数据大多具有空间属性^[2],不是单纯的数字变化。空间回归模型中,空间统计数据的空间多维特征和时空相关的假设,这两点是区别传统统计的最大特征,依据空间数据能了解研究对象的各种方位的详细特性,以此不仅使得很多疾病的研究得以有更好的模型,而且更能揭示疾病的影响因素及其时空分布。

2. 常用的空间回归模型:在进行回归模型拟合之前,需要查看数据的空间自相关性质。空间自相关包括两种类型:全局空间自相关和局部空间自相关。全局空间自相关性,用Moran'I来表示,用于研究某一属性取值在整个空间上的空间聚集状态。将全局空间自相关性分析的Moran's I系数分解到局域空间上,Anselin^[14]提出了局部空间自相关性指示变量(LISA),研究空间单元属性取值在某些局域位置上的空间分布状态^[15]。全局自相关将整个区域相关性进行检验,用于根据背景知识了解区域变化有无相似趋势,而局域自相关则是了解每个小区域的聚集情况存在的差异。除了空间自相关性的检验,还应注意空间回归模型对邻近单位的定义,即邻近单位的选择是根据权重矩阵生成,一般有三种方式可选择:①所有与选定单位共有—个边界的单位^[16];②所有与选定单位的距离在规定的阈值距离之内的单位;③所有与选定单位距离最近的单位^[17]。

(1)空间滞后模型(spatial lag model, SLM):由Anselin提出的空间自回归模型一般形式为^[18], $y = \rho W_1 y + X\beta + e$, $e = \lambda W_2 e + \mu$, $\mu \sim N(0, \sigma^2)$ 。式中 y 为 $n \times 1$ 应变量, X 为 $n \times 1$, β 为 $k \times 1$ 回归系数, e 为残差项, ρ 为空间相关系数, λ 为残差相关系数, μ 为随机误差项并服从正态分布, W_1 、 W_2 为空间权重矩阵。当 $\rho \neq 0, \beta \neq 0, \lambda = 0$ 时,空间自回归模型转换为SLM: $y = \rho W_1 y + X\beta + \mu$, $\mu \sim N(0, \sigma^2)$ 。

各系数意义同空间自回归模型一般形式,在不同领域的不同数据,拟合出的模型可对变量各有解释。SLM适合估计是否存在空间相互作用以及空间相互作用的强度,可用于探讨各变量在某地区是否具有扩散现象。

(2)空间误差模型(spatial error model, SEM)^[18]:空间自

DOI: 10.3760/cma.j.issn.0254-6450.2013.011.024

基金项目:国家自然科学基金(30872159);国家科技重大专项(2012ZX10004-220)

作者单位:200032 上海,复旦大学公共卫生学院流行病学教研室 公共卫生安全教育部重点实验室 复旦大学热带病研究中心

通信作者:周艺彪, Email: ybzhou@fudan.edu.cn

回归模型一般形式中,当 $\rho=0, \beta \neq 0, \lambda \neq 0$ 时,转换为 SEM, $y = \lambda W_2 e + X\beta + \mu, \mu \sim N(0, \sigma^2)$ 。

各系数解释同空间自回归模型一般形式。SEM 是误差项在空间相关时的模型,地区间的相互关系通过误差项来体现,用于度量邻近地区关于因变量的误差对本地区观察值的影响程度。

(3)地理加权回归模型(geographical weighted regression, GWR):由 Nakaya 等^[19]将基于空间位置的关系矩阵加到经典回归分析模型中而形成,其经典回归方程: $Y_i = \alpha_0 + \sum \alpha_k K_{ik} + \varepsilon_i$ 。式中 K 为变量数, i 为样本号, α_0 为初始参数, ε_i 为第 i 个空间点误差,相应的系数估计 $\alpha = (X^T X)^{-1} X^T Y$ 。GWR 方程为 $Y_i = \alpha_0(u_i, v_i) + \sum \alpha_k(u_i, v_i) K_{ik} + \varepsilon_i$ 。式中 (u_i, v_i) 为第 i 格中心点坐标。GWR 的回归系数 β 不再是全局性的统一单值,而是随空间位置改变而变化的模型参数,并用以探索空间数据的空间异质性^[20]。对于局域性空间变化,数据在空间上不是同一变化规律,在不同区域间自变量对因变量的影响不同,可采用 GWR。该模型中,特定区位的回归系数不再是利用全部信息获得的假定常数,而是利用邻近观测值的子样本数据进行局域回归估计,并随着空间局域地理位置变化而变化的变数。

3. 空间回归模型在公共卫生方面的应用:

(1)国外应用情况:Weidmann 等^[21]通过拟合 SLM,发现德国献血人群有明显的地区聚集性,揭示了献血参与程度与该地区人群年龄等因素有负相关性,为政府提供了献血区域的选择策略。Chen 和 Wen^[15]等同样应用该模型研究了中国台湾地区肥胖危险因子的空间关系,发现 2001—2005 年肥胖危险因素发生很大变化,并从模型中反映某些地区存在很强的空间溢出效应,提示在这些地区需要政府政策干预。Semaan 等^[22]以州为单位,分析美国各州人群经济状况与患性传播疾病(STD)间的关系,分别使用简单最小二乘法和 SLM 进行分析,结果二者参数估计相似。即如不考虑人种构成,高经济水平地区的 STD 感染率低;如考虑人种构成,则美籍非裔人群多的地区其 STD 感染率高,但此时经济水平与 STD 感染率无关。Yu 等^[23]采用 SLM 研究美国新泽西州人口学因素与香烟销售点密度间的关系,结果表明烟草商更倾向于在经济条件较差的地区增加销售点,在新泽西州西班牙裔为购买烟草最多的人群,其次为中产收入者和美籍黑人。Duncan 等^[24]采用该模型分析了建筑环境对青年抑郁症的影响,结合空间属性的分析结果显示,建筑环境对抑郁症的影响很小。上述研究采用 SLM 均考虑了被研究因素有很强的空间自相关性,可以估计出空间相互作用的强度,并可探讨区域溢出效应。

国外研究中多见于将多种空间回归模型进行拟合再选择。Conley^[17]应用 SLM 和 SEM 分析美国有毒物质排放清单与肺癌死亡率间的空间关系。Yang 和 Jin^[25]应用空间回归方法,分析和预测美国锡达河流域特点对 NO_3 、 NO_2 浓度的影响,发现空间回归方法比普通最小二乘法更适用于污染源非点分布的情况。Hu 和 Rao^[26]同样采用两种模型,分析评价美国东部地区气溶胶光学厚度(AOD)与慢性缺血性心脏病(CIHD)间的关系,发现该地区两者间存在空间聚集性,拟合模型提示 AOD 是 CIHD 的危险因素。Owusu-Edusei 和 Chesson^[27]分别

采用多种空间回归模型拟合,研究了德克萨斯州以县为单位衣原体和淋病与人种间的关系,通过指标选择,得出空间误差模型最为符合数据情况。上述研究各因素存在强的自相关性,同时误差项在空间上也有相关性,可采用 SLM 和 SEM 分别拟合,再根据检验指标选择更为合适的模型。

(2)国内应用情况:戴平生和李芳芳^[28]分析全国 31 个省市城乡居民医疗保险消费支出的影响因素,首先使用一般线性模型,得到人均可支配收入对医疗保健支出为负影响,违反凯恩斯消费函数理论,后经检验数据存在空间自相关性,故采用 SEM,分析表明住院医药费、地方财政卫生支出、受教育年限是影响城市医疗消费的主要因素,而个人纯收入、地方财政卫生支出、受教育年限是影响农村医疗消费的主要因素。黄秋兰等^[29]在全局和局部两水平上采用 SLM,分析广西地区流行性乙型脑炎与气象因素间的关系,结果显示该地区年均相对湿度、年均日照、年均气压是影响发病的主要气象因素,且具有空间变异性。陈炳为等^[29]使用 GWR 发现参数估计在各区域并非完全相等,甲状腺肿大率较高的区域主要位于四川盆地中部及南部山区和丘陵地带。李晓和薛付忠^[30]研究发现采用 GWR 能够准确描述人群空间遗传结构与气候因素间的空间变化,较全局模型具有明显的优越性。刘云霞等^[31]利用该模型探讨山东省结核病及其影响因素间的局域关系,为制定适宜的结核病防控策略提供依据。上述研究中,无法将整个地区的数据进行分析,需先划分为若干个小区域分别进行模型拟合,适宜采用 GWR 方法。

程娟^[10]对传染性非典型肺炎(SARS)、婴儿猝死综合征(SIDS)、婴儿肾结石采用空间回归分析。提示地区人口总数、医院总数与 SARS 传播关系无关,而地区人口密度、医师总人数与 SRAS 传播密切相关;SIDS 与种族因素有关;婴儿肾结石病例数与农村人口比例、人均用水量和水污染情况存在较强相关性。唐启强等^[32]采用 SLM 和 SEM 分析鄱阳湖区南昌县血吸虫疫情分布特点和流行因素,结果表明疫水接触强度、村内耕牛数量、村内湖岸线长度、村到鄱阳湖边界的平均距离与血吸虫病患者数具有明显的相关性,其中患者例数和行政村到鄱阳湖沿岸的距离呈负相关。张洪梅^[33]采用两种模型拟合分析山东省肺结核病空间分布及其影响因素,结果表明结核病发病与该地区的生活方式、生活习惯、人口密度、人口流动速率及生产力水平等因素高度相关。尼楚君和王怀明^[34]采用 SLM 和 SEM 分析了经济收入差距对农村居民健康影响的空间效应,结果显示我国 30 个省区市农村居民健康具有显著的空间相关性,并具有空间溢出效应和空间误差冲击效应,各省区市农村居民经济收入差距对其健康影响同样具有空间溢出效应和空间误差冲击效应。解丕^[35]同样使用两种模型分析 1997—2004 年 30 个省区市政府卫生投资的空间特征,表明政府卫生投资具有空间溢出作用并逐年增强,政府卫生投资在空间上存在显著的互补效应。

4. 小结:公共卫生领域诸多问题存在着连续性和区域性和聚集性等特点,空间回归模型能够较好地结合这些特点进行分析,从空间环境角度观察其发生和流行的时空传播规律,从而发现分布特点、空间聚集性、空间分布的差异,并发

现这些差异与聚集性的原因,为预防和控制公共卫生领域中的问题有着很大的帮助。

参 考 文 献

- [1] Zhang KJ. Research and application of spatial autoregressive models. Nanning: Guangxi University, 2012. (in Chinese)
张可娟. 空间自回归模型的研究与应用. 南宁: 广西大学, 2012.
- [2] Use of geographic information system in epidemiology (GIS-Epi). *Epidemiol Bull*, 1996, 17(1): 1-6.
- [3] Yang KZ, Feng DT, Shen TY. The latest research progress of spatial econometrics. *Res Deve*, 2009(2): 7-12. (in Chinese)
杨开忠, 冯等田, 沈体雁. 空间计量经济学研究的最新进展. 开发研究, 2009(2): 7-12.
- [4] LA. *Spatial econometrics: methods and models*. Dordrecht: Kluwer Academic, 1988.
- [5] Fischer M, Getis A. Recent developments and in spatial analysis: spatial statistics, behavioral modelling and computational intelligence. 2002.
- [6] Paciorek CJ. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Stat Sci*, 2010, 25(1): 107-125.
- [7] Perez SI, Diniz-Filho JA, Bernal V, et al. Spatial regression techniques for inter-population data: studying the relationships between morphological and environmental variation. *J Evol Biol*, 2010, 23(2): 237-248.
- [8] Tang QQ. Evolution of spatial-temporal pattern of schistosomiasis in Nanchang county. Nanchang: Jiangxi Normal University, 2012. (in Chinese)
唐启强. 南昌县血吸虫疫情时空格局演化研究. 南昌: 江西师范大学, 2012.
- [9] Buck T, Plicht B, Kahlert P, et al. Understanding the asymmetrical Vena Contracta Area: the difficult relationship between 2D and 3D measurements. *JACC Cardiovasc Imaging*, 2013, 6(6): 744.
- [10] Cheng J. *Spatial statistics application for medicine*. Wuhan: Huazhong University of Science and Technology, 2009. (in Chinese)
程娟. 空间统计在医学中的应用. 武汉: 华中科技大学, 2009.
- [11] Gao HW. Prediction of malaria transmission risk based on environmental factors of small-scale spatial and temporal patterns. Beijing: Chinese People's Liberation Army Military Medical Science Academy, 2012. (in Chinese)
高宏伟. 不同时空尺度疟疾传播风险预测及环境影响因素研究. 北京: 中国人民解放军军事医学科学院, 2012.
- [12] Liu XD. A study on the spatial and temporal distribution of HFRS in China and the impact of climate factors on HFRS in Liaoning. Jinan: Shandong University, 2012. (in Chinese)
刘晓冬. 中国肾综合征出血热时空分布及气候因素对辽宁省HFRS影响的研究. 济南: 山东大学, 2012.
- [13] Cheng CL, Chen YC, Liu TM, et al. Using spatial analysis to demonstrate the heterogeneity of the cardiovascular drug-prescribing pattern in Taiwan. *BMC Public Health*, 2011, 11: 380.
- [14] Anselin L. Local indicators of spatial association—LISA. *Geograph Anal*, 1995, 2(27): 93-115.
- [15] Chen DR, Wen TH. Elucidating the changing socio-spatial dynamics of neighborhood effects on adult obesity risk in Taiwan from 2001 to 2005. *Health Place*, 2010, 16(6): 1248-1258.
- [16] Zhang SL, Zhang K. Grid-based spatial weight matrix of discrete points. *J Geod Geodynam*, 2008, 28(6): 87-90. (in Chinese)
张松林, 张昆. 用格网法建立离散点的空间权重矩阵方法研究. 大地测量与地球动力学, 2008, 28(6): 87-90.
- [17] Conley JF. Estimation of exposure to toxic releases using spatial interaction modeling. *Int J Health Geogr*, 2011, 10: 20.
- [18] Xu BY, Chen BW, Ni ZZ, et al. Analysis on variation of urinary iodine of children by spatial autoregressive model. *J Hyg Res*, 2004(5): 578-580. (in Chinese)
许碧云, 陈炳为, 倪宗瓚, 等. 利用空间自回归模型分析儿童尿碘的变异. 卫生研究, 2004(5): 578-580.
- [19] Nakaya T, Fotheringham AS, Brunson C, et al. Geographically weighted Poisson regression for disease association mapping. *Stat Med*, 2005, 24(17): 2695-2717.
- [20] Huang QL, Tang XY, Zhou HX, et al. Study on the relationship between Japanese encephalitis and meteorological factors in Guangxi, based on global and local spatial regression model. *Chin J Dis Control Prev*, 2013, 17(4): 282-286. (in Chinese)
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.
- [21] Weidmann C, Schneider S, Litaker D, et al. A spatial regression analysis of German community characteristics associated with voluntary non-remunerated blood donor rates. *Vox Sang*, 2012, 102(1): 47-54.
- [22] Semaan S, Sternberg M, Zaidi A, et al. Social capital and rates of gonorrhoea and syphilis in the United States: spatial regression analyses of state-level associations. *Soc Sci Med*, 2007, 64(11): 2324-2341.
- [23] Yu D, Peterson NA, Sheffer MA, et al. Tobacco outlet density and demographics: analysing the relationships with a spatial regression approach. *Public Health*, 2010, 124(7): 412-416.
- [24] Duncan DT, Piras G, Dunn EC, et al. The built environment and depressive symptoms among urban youth: a spatial regression study. *Spat Spatiotemporal Epidemiol*, 2013, 5: 11-25.
- [25] Yang X, Jin W. GIS-based spatial regression and prediction of water quality in river networks: a case study in Iowa. *J Environ Manage*, 2010, 91(10): 1943-1951.
- [26] Hu Z, Rao KR. Particulate air pollution and chronic ischemic heart disease in the eastern United States: a county level ecological study using satellite aerosol data. *Environ Health*, 2009, 8: 26.
- [27] Owusu-Edusei KJ, Chesson HW. Using spatial regression methods to examine the association between county-level racial/ethnic composition and reported cases of Chlamydia and gonorrhoea: an illustration with data from the state of Texas. *Sex Transm Dis*, 2009, 36(10): 657-664.
- [28] Dai PS, Li FF. Outcome prediction of mild cognitive impairment based on multi-state Markov model. *Chin J Health Stat*, 2012(4): 514-515. (in Chinese)
戴平生, 李芳芳. 基于误差空间模型城乡居民医疗保健消费影响因素的实证分析. 中国卫生统计, 2012(4): 514-515.
- [29] Chen BW, Xu BY, Ni ZZ, et al. The application of geographical weighted models to prevalence of endemic goiter. *Appl Stat Manage*, 2005(3): 113-117. (in Chinese)
陈炳为, 许碧云, 倪宗瓚, 等. 地理权重回归模型在甲状腺肿大中的应用. 数理统计与管理, 2005(3): 113-117.
- [30] Li X, Xue FZ. Application of the geographically weighted regression model in spatial genetic structure of the human population. *J Shandong Univ: Health Sci*, 2011(2): 119-124. (in Chinese)
李骁, 薛付忠. 地理权重回归在人类群体空间遗传结构中的应用. 山东大学学报: 医学版, 2011(2): 119-124.
- [31] Liu YX, Liu YX, Zhang BB, et al. Spatial epidemiology study on tuberculosis based on geographical weighted regression model. *Chin J Antituberc*, 2013(5): 343-346. (in Chinese)
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(5): 343-346.
- [32] Tang QQ, Zhang Z, Zhao A, et al. Spatial analysis on epidemic factors of schistosomiasis in Nanchang county over the coast of Poyang Lake. *Trop Geograph*, 2013(1): 76-80. (in Chinese)
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(1): 76-80.
- [33] Zhang HM. Study on spatial distribution of tuberculosis and influential factors in Shandong province. Jinan: Shandong University, 2009. (in Chinese)
张洪梅. 山东省肺结核病的空间分布及其影响因素研究. 济南: 山东大学, 2009.
- [34] Ni CJ, Wang HM. Spatial effect analysis of income gap of rural residents' health. *Quest*, 2011(4): 1-4. (in Chinese)
尼楚君, 王怀明. 收入差距对农村居民健康影响的空间效应分析. 求索, 2011(4): 1-4.
- [35] Xie E. Spatial effect of public health expenditure: substitution or complement. *Chin Health Econom*, 2008(4): 24-27. (in Chinese)
解垂. 政府卫生投资的空间作用: 互补抑或替代. 中国卫生经济, 2008(4): 24-27.

(收稿日期: 2013-06-23)

(本文编辑: 张林东)