

使用反距离权重内插法绘制中国 1990 年代肿瘤分布地图

曲宸绪 姜勇 武燕萍 邹小农 乔友林

【摘要】 目的 绘制中国 1990 年代死因回顾 1/10 人口抽样调查肿瘤分布地图,展示中国 1990 年代肿瘤的地理分布和区域性集中趋势。方法 利用 1990 年代的全国死因抽样调查数据,根据中国 1964 年人口年龄分布计算标准化死亡率,以中国 1990 年代数字地图为背景,在 Arcview 软件支持下,利用反距离权重内插法(IDW)预测非采样地区的死亡率数值,绘制肿瘤死亡分布地图。结果 现已完成全国 1990 年代肿瘤分布地图绘制,与填充法相比,IDW 所绘地图死亡率分布边缘且过渡平滑,显示出肿瘤地区分布特点。结论 利用 IDW 根据抽样区域数据预测其他区域数据,可以实现用抽样数据绘制全国肿瘤分布地图,结果仍可以较为准确地反映出肿瘤分布规律。

【关键词】 肿瘤;地理信息系统;反距离权重内插法

Study on the development of a choropleth atlas on cancer mortality using the inverse distance weight interpolation in the 1990's QU Chen-xu, JIANG Yong, WU Yan-ping, ZOU Xiao-nong, QIAO You-lin. Department of Epidemiology, Cancer Institute and Cancer Hospital Chinese Academy of Medical Science, Peking Union Medical College, Beijing 100021, China
Corresponding author: QIAO You-lin, Email: qiaoy@public.bta.net.cn

【Abstract】 Objective To draw a chorochromatic atlas of mortality on China Cancer Database in order to display the geographical distribution of selected diseases in China and to identify its demographic and disease-specific patterns in the 1990's. **Methods** The source of data was from nationwide cause-of-death surveys conducted in the 1990's. Standardized rates were computed by direct method using the population age distribution in 1964 as the standard of weights. Inverse distance weight(IDW) was applied as the method of interpolation with the help of Arcview system to draw a choropleth map of cancer mortality. **Results** The IDW maps of cancer mortality shows a continuous and smooth variation, especially compared with maps drawn by filling method. **Conclusion** With the application of inverse distance weight interpolation, it seemed feasible to draw continuous map of cancer on sampling data.

【Key words】 Carcinoma; Geographic information system; Inverse distance weight

卫生部全国肿瘤防治办公室 1973 - 1975 年曾进行一次“全国人口三年肿瘤死亡回顾调查”,并根据调查数据用填充法绘制了《中华人民共和国恶性肿瘤地图集》,该地图集首次揭示了国内肿瘤的分布规律,推动和促进了以后几十年肿瘤流行病学研究,对我国肿瘤防治工作具有重要意义。1990 - 1992 年,国家又组织了一次全国性的死亡回顾调查,目的是评测经过近 20 年人民生活水平明显提高情况下,

国内肿瘤的分布规律和特点,考虑项目经费原因及成本效益,回顾调查按 1/10 人口比例随机抽样进行。因为是抽样数据,空间分析(spatial analysis)法没有得到推广,所以一直未绘制出我国 1990 年代的肿瘤分布地图集。本文主要论述利用空间分析技术,绘制我国 1990 年代肿瘤分布地图,完善肿瘤死因资料,为肿瘤危害的评估和防治提供更多依据。

资料与方法

1. 死亡率数据:采用卫生部全国肿瘤防治办公室 1990 - 1992 年所组织的全国人口死亡原因抽样调查数据,共涉及除台湾、西藏、青海和新疆以外的 27 个省、自治区、直辖市随机样本地区人群,共调查

基金项目:科技部专项课题基金资助项目(2000DEB30104)

作者单位:100021 北京,中国医学科学院肿瘤医院/肿瘤研究所流行病学研究室

通讯作者:乔友林, Email: qiaoy@public.bta.net.cn

263 个县(市),按我国 1964 年人口分布计算各肿瘤的年龄别标准化死亡率。

2. 中国数字地图:采用“美国国际地球科学信息网络中心”(CIESIN)所发布的 1990 年 7 月 1 日中国省级和县级行政区域 $1:10^6$ 地理数据,由美国哥伦比亚大学社会经济数据与应用中心、中国测绘科学研究院及华盛顿大学“中国时空计划”协作开发。

3. 反距离权重内插法(inverse distance weight, IDW):基本公式

$$F(x, y) = \sum_{i=1}^n W_i f_i$$

$$W_i = \frac{H_i^{-p}}{\sum_{j=1}^n H_j^{-p}}$$

$$H_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$$

式中, n : 某一区域内样本点的数目; f_i : 区域 i 的观测值; W_i : 被赋与的权重; (x, y) 预测点的座标; (x_i, y_i) 样本点的座标; H_i : 预测点与样本点的距离; p : 方次参数, 一个正实数, 用来调整权重与距离的关系, 通常取 2。

IDW 是一种加权平均内插法, 该方法认为任何一个观测值都对邻近的区域有影响, 且影响的大小随距离的增大而减小。方次参数控制着权重系数如何随着离开一个格网结点距离的增加而下降。对于一个较大的方次, 较近的数据点被给定一个较高的权重份额, 对于一个较小的方次, 权重比较均匀地分配给各数据点。计算一个格网结点时给予一个特定数据点的权值与指定方次的从结点到观测点的该结点被赋予距离倒数成比例。当计算一个格网结点时, 配给的权重是一个分数, 所有权重的总和等于 1.0。当一个观测点与一个格网结点重合时, 该观测点被给予一个实际为 1.0 的权重, 所有其他观测点被给予一个几乎为 0.0 的权重。

4. 方法: 在 Arcview 3.2 软件中, 利用其空间运算功能, 计算全国各县的中心点(center); 以点数据形式输入各地区 1990 年代各肿瘤死亡率, 建立点矢量图, 并按百分位数法分为 6 等级进行着色; 应用空间分析(spatial analysis 2.0a)模块的空间预测功能, 以临近 6 个点为基础, 利用 IDW 建立我国 1990 年代全国肿瘤死因分布矢量图; 绘制等高线、图例等, 导出肿瘤分布地图; 利用空间分析模块的山体阴影功能, 绘制 IDW 图层的山体阴影(hillshade), 将肿瘤死亡分布矢量图的着色方法更换为山体阴影, 绘制全国肿瘤死因分布三维分布图。

结 果

现以我国 1990 年代男性食管肿瘤及鼻咽肿瘤为例说明 IDW 制图结果。我国的食管癌有明显的地区分布规律, 高发区与低发区发病率和死亡率相差几十倍, 按常用的填充法绘制 1990 年代男性食管癌肿瘤分布地图, 由于大部分地区缺少数据, 只能得到散在的图斑(图 1), 尽管可反映出抽样地区的死亡率高, 但很难看出该肿瘤的地区分布规律。采用 IDW 绘制的 1990 年代男性食管癌地区分布预测地图填补了空白区域, 高发区、低发区显示完整, 地图所示与我国食管癌分布特点基本一致, 可清楚看到高发区位于在河南、河北、山西三省交界^[1,2], 江苏省北部及部分沿海地区^[3]。图 3 用 Z 值表示死亡率的高低, 有一定立体感, 相比平面的地图更加直观。

鼻咽癌在我国主要集中于广东省, 由图 4 可以看出, 广东省抽样地区的鼻咽癌死亡率明显高于其他地区, 但高发范围却难以确认, 图 5、6 采用 IDW 绘制, 借助等高线可较为清晰的看出鼻咽癌高发区域为广东省中部, 广西壮族自治区南部及海南省南部次之。

讨 论

流行病学研究中, 经常要利用样本预测总体性质, 但所用方法一般针对一维属性的数据, 对于具有二维或多维空间属性数据的预测, 由于所涉及的数学知识在流行病学研究中运用较少, 且相关计算量大, 并不为多数流行病学研究者所掌握。近年来地理信息系统(GIS)技术的发展带动了一系列相关学科, 其中包括空间统计学的发展和成熟, 对于空间数据预测现已有多套方案。而且随信息技术的发展, GIS 应用得到推广, 为流行病学研究提供了方便的制图及空间分析工具^[4], 对于流行病学研究者来说, 只要具备相关知识, 进行空间分析运算已不再困难。

根据邻近区域属性值有较大相似性^[5,6], 在某一限定空间内利用已知区域属性值推测未知区域属性的过程, 在空间分析中称为内插(插值)(interpolation)。内插法常用于将离散点的测量数据转换为连续的数据曲面, 以便与其他空间现象的分布模式进行比较。IDW 是目前一种广泛使用的内插法, 运算简便, 适用范围广, 是一种准确预测的

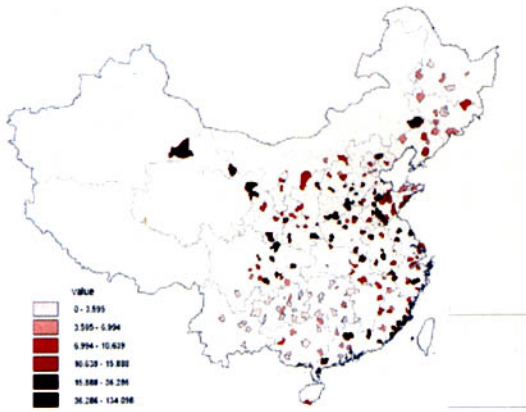


图1 我国 1990 年代男性食管肿瘤地区分布(填充法)

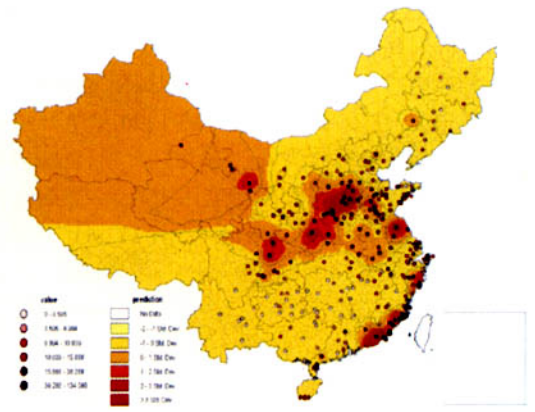


图2 我国 1990 年代男性食管肿瘤死亡分布预测(IDW)

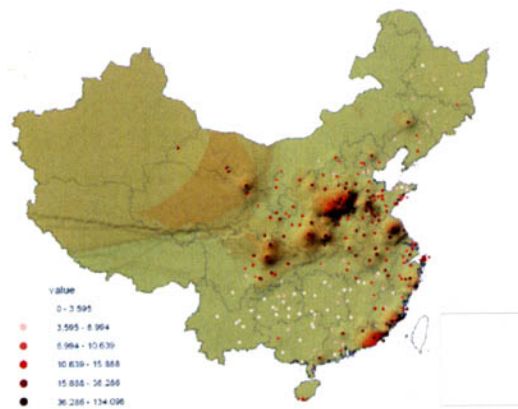


图3 我国 1990 年代男性食管肿瘤死亡率三维分布(IDW)

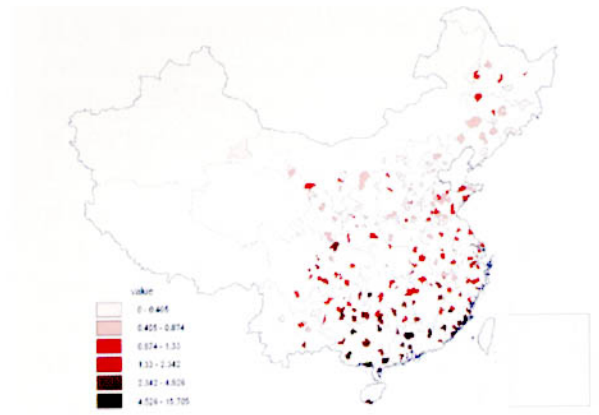


图4 我国 1990 年代男性鼻咽肿瘤地区分布(填充法)

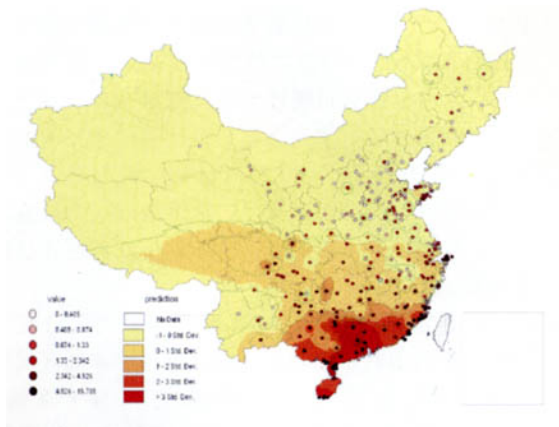


图5 我国 1990 年代男性鼻咽肿瘤死亡分布预测(IDW)

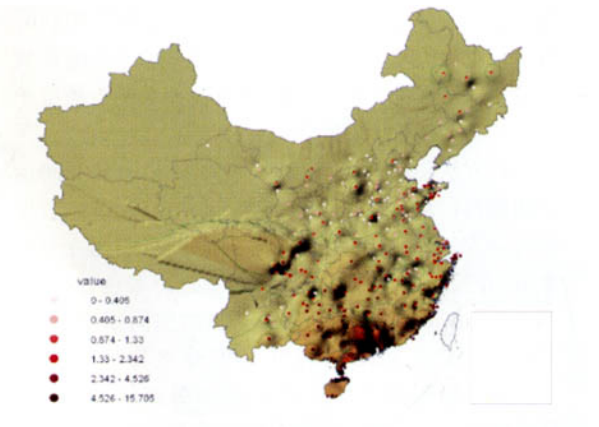


图6 我国 1990 年代男性鼻咽肿瘤死亡率三维分布(IDW)

算法模型,即预测值构成的曲面严格通过每一个观测值构成顶点(图 6),在进行复杂的拟合运算前常用来获得对抽样数据的初步印象。另一方面 IDW 预测值分布与观测值分布十分接近,我国 1990 年代男性食管癌死亡率调查值与预测值(除新疆、青海、西藏、台湾四省区)的统计特点,均值、标准差、偏度系数与峰度系数等统计指标十分接近(表 1)。由于观测值的随机波动, IDW 计算的数据曲面一般不够平滑(smooth)。其特征之一是要在格网区域内产生围绕观测点位置的“牛眼”(即与观测点重合的网格预测值和周边值有差异)。图 2 与图 3 可以较清楚的看到四川省北部及陕西省与河南省交界处的“牛眼”,不过由于绘制 1990 年代肿瘤分布地图并不要求准确的预测值,且按一倍标准差分等级着色,所以“牛眼”现象并不突出。除 IDW 之外,还有几种内插法可以采用,如广域多项式(global polynomial)制图平滑,克里格(Kriging)法能够评估预测值的标准误差及可能性大小^[7,8],一般参数较多,算法灵活,可根据数据性质及研究目的灵活运用,常作为 IDW 的后继分析方法。

表1 我国 1990 年代男性食管癌观测值与观测值统计

项目	观测值	预测值
例数	340	20844
最小值	0	0.08
最大值	134.10	133.80
均值	16.53	14.63
标准差	19.61	14.60
方差	384.45	213.07
偏度系数	2.85	2.27
峰度系数	10.44	7.16

绘制我国 1990 年代肿瘤死亡分布地图的目的是揭示全国肿瘤分布规律,理想的肿瘤分布地图应校正随机因素的影响,完整体现高发区与低发区的分布,并且过渡平滑。IDW 不足之处在于没有对观测数据的波动性进行校正,尽管数据来源于 3 年的死因回顾查,相对其他一些短期的流行病学调查来说,时间跨度大,观测值的稳定性较好,但对于人口较少区县及一些相对罕见的肿瘤来说,所受随机因

素的影响仍难以忽视^[9],而年龄别标化对于这种随机波动没有校正作用。在 1990 年代的肿瘤分布地图中,偶尔可以发现这种过高或过低的异常值。尽管目前已有多种方法校正异常值,但校正方法的选择一般涉及多种因素,如肿瘤分布规律、肿瘤的种类及所在地区,一般来说很难找出一种所有人都认为合理的校正方法。IDW 本身就包括一定的校正,但由于 IDW 的权重主要取决于预测点与观测点的距离,涉及的因素较为简单,必然导致一些影响因素被忽视。但总的来说,目前简单与复杂的内插法对预测的总体趋势较为接近,当研究并不追求预测的准确性,如这次绘制的我国 1990 年代地图,目的在于反映出肿瘤地理分布特征,而不是要利用预测数据进一步分析,这种情况下 IDW 不失为一种好的方法,制图结果也较为可信。

参 考 文 献

- 1 宋建荣,周有尚. 食管癌地理环境因素的流行病学分析. 中华流行病学杂志,1992,13:329-332.
- 2 钱宇平. 流行病学进展. 第 4 卷. 北京:人民卫生出版社,1984. 174-298.
- 3 饶克勤,李连弟. 中国恶性肿瘤危险因素研究. 第 1 版. 北京:中国协和医科大学出版社,2003. 235-240.
- 4 Patrick WO, William AY, Elizabeth M. Public health informatics and information system. 1st ed. New York:Springer-Verlag, 2003. 431-467.
- 5 Harvey JM. Tobler's first law and spatial analysis. Ann Assoc Am Geographers,2004,94:284-289.
- 6 Kelsall J, Wakefield J. Modelling spatial variation in disease risk: a geostatistical approach. J Am Sta Assoc,2002, 97:692-701.
- 7 Carrat F, Valleron AJ. Epidemiologic mapping using the "kriging" method: Application to an influenza-like epidemic in France. Am J Epidemiol,1992,135:1293-1300.
- 8 Webster R, Oliver MA, Muir KR, et al. Kriging the local risk of a rare disease from a register of diagnoses. Geographical Analysis, 1994, 26:168-185.
- 9 Oliver MA, Muir KR, Webster R, et al. A geostatistical approach to the analysis of pattern in rare disease. J Public Health Med,1992, 14:280-289.

(收稿日期:2005-10-13)

(本文编辑:张林东)