

中国乳腺癌发病率空间分布数据集（2014–2016）

王若菡^{1,2}, 王培涵^{1,2}, 徐成东^{1,2}, 王伟^{1,2}, 王振波^{1,2*}

1. 中国科学院地理科学与资源研究所, 北京 100101; 2. 中国科学院大学, 北京 100049

摘要: 作者利用 2017–2019《中国肿瘤登记年报》统计, 对其收录的 2014–2016 年全国各种肿瘤登记处乳腺癌发病率进行整编, 通过 ArcGIS10.8 软件对中国县级市乳腺癌发病率进行矢量化和空间可视化。作者使用描述性统计与空间统计分析作为研究方法, 探究中国乳腺癌发病率的区域差异、空间分布和变化趋势, 得到 2014–2016 年中国县级乳腺癌发病率分布数据集。该数据集内容包括中国 2014–2016 年下列数据: (1) 县级乳腺癌发病率数据; (2) 县级乳腺癌发病率分组统计; (3) 东中西部乳腺癌发病率统计。数据集存储为.shp 和.xlsx 格式, 由 25 个数据文件组成, 数据量为 21.5 MB。

关键词: 中国; 乳腺癌; 发病率; 空间分布

DOI: <https://doi.org/10.3974/geodp.2024.02.04>

CSTR: <https://cstr.escience.org.cn/CSTR:20146.14.2024.02.04>

数据可用性声明:

本文关联实体数据集已在《全球变化数据仓储电子杂志（中英文）》出版, 可获取:

<https://doi.org/10.3974/geodb.2024.06.06.V1> 或 <https://cstr.escience.org.cn/CSTR:20146.11.2024.06.06.V1>.

1 前言

乳腺癌是目前全球最常见的癌症之一, 其发病率在北美、欧洲和澳大利亚等发达地区最高^[1]。中国乳腺癌发病率在世界范围内处于较低水平, 然而在过去三十年中, 发病率增加了 20%–30%, 每年增长约 3%–5%, 高于全球 1.5% 的平均增长率^[2]。中国国家癌症中心 (NCC) 发布的全国癌症统计数据表明, 在中国女性中, 乳腺癌是发病率最高的癌症^[3]。随着中国经济社会的快速发展、人口数量增长和老龄化趋势, 以及主要危险因素的流行, 乳腺癌的负担在未来可能会更加严峻^[4]。因此, 迫切需要对中国乳腺癌进行相关研究, 以应对和控制由乳腺癌风险带来的个人和社会经济负担。

中国国土面积广阔, 其社会经济和自然环境在不同区域之间存在着显著差异, 如果简单的将中国视为一个整体, 就会忽视风险因素对疾病影响的区域异质性, 从而可能导致研究结论与实际情况不符^[5]。目前关于中国乳腺癌空间分布及其影响因素的研究大多局限于

收稿日期: 2024-01-12; 修订日期: 2024-05-10; 出版日期: 2024-06-25

基金项目: 国家自然科学基金 (42130713); 中华人民共和国科学技术部 (2019QZKK1005)

*通讯作者: 王振波, 中国科学院地理科学与资源研究所, wangzb@igsrr.ac.cn

数据引用方式: [1] 王若菡, 王培涵, 徐成东等. 中国乳腺癌发病率空间分布数据集 (2014–2016) [J]. 全球变化数据学报, 2024, 8(2): 147–154. <https://doi.org/10.3974/geodp.2024.02.04>. <https://cstr.escience.org.cn/CSTR:20146.14.2024.02.04>.

[2] 王若菡, 王培涵, 徐成东等. 中国乳腺癌发病率空间分布数据集 (2014–2016) [J/DB/OL]. 全球变化数据仓储电子杂志, 2024. <https://doi.org/10.3974/geodb.2024.06.06.V1>. <https://cstr.escience.org.cn/CSTR:20146.11.2024.06.06.V1>.

单个城市或省份等小区域^[6-9]，而在中国进行全国范围的研究相对较少，且主要以省级作为研究尺度^[10-12]。空间尺度的局限可能导致研究结果缺乏精确性，难以全面地反映出中国乳腺癌的真实情况。

本数据集以县级为研究单位，对 2017–2019《中国肿瘤登记年报》各肿瘤登记处的乳腺癌发病率数据进行整编^[13-15]，在此基础上，通过 ArcGIS10.8 软件转换为矢量数据。通过图表统计、矢量数据可视化和空间自相关分析，探究中国乳腺癌发病率的区域差异、空间分布及变化趋势，为制定针对性的预防和控制策略提供科学依据。

2 数据集元数据简介

《中国乳腺癌发病率空间分布数据集（2014–2016）》^[16]的名称、作者、地理区域、数据年代、数据集组成、数据出版与共享服务平台、数据共享政策等信息见表 1。

表 1 《中国乳腺癌发病率空间分布数据集（2014）》元数据简表

条目	描述
数据集名称	中国乳腺癌发病率空间分布数据集（2014–2016）
数据集短名	BreastCancerIR2014-2016
作者信息	王若茜，中国科学院地理科学与资源研究所，wangruohan2446@igsnrr.ac.cn 王培涵，中国科学院地理科学与资源研究所，wph1996@126.com 徐成东，中国科学院地理科学与资源研究所，xucd@lreis.ac.cn 王伟，中国科学院地理科学与资源研究所，wang_wei@lreis.ac.cn 王振波，中国科学院地理科学与资源研究所，wangzb@igsnrr.ac.cn
地理区域	中国
数据年代	2014–2016 年
数据格式	.shp、.xlsx
数据量	21.5 MB
数据集组成	2014–2016 年中国乳腺癌发病率空间分布矢量数据、表格统计数据
基金项目	国家自然科学基金（42130713）；中华人民共和国科学技术部（2019QZKK1005）
出版与共享服务平台	全球变化科学研究数据出版系统 http://www.geodoi.ac.cn
地址	北京市朝阳区大屯路甲 11 号 100101，中国科学院地理科学与资源研究所
数据共享政策	（1）“数据”以最便利的方式通过互联网系统免费向全社会开放，用户免费浏览、免费下载；（2）最终用户使用“数据”需要按照引用格式在参考文献或适当的位置标注数据来源；（3）增值服务用户或以任何形式散发和传播（包括通过计算机服务器）“数据”的用户需要与《全球变化数据学报（中英文）》编辑部签署书面协议，获得许可；（4）摘取“数据”中的部分记录创作新数据的作者需要遵循 10%引用原则，即从本数据集中摘取的数据记录少于新数据集总记录量的 10%，同时需要对摘取的数据记录标注数据来源 ^[17]
数据和论文检索系统	DOI, CSTR, Crossref, DCI, CSCD, CNKI, SciEngine, WDS, GEOSS, PubScholar, CKRSC

3 数据研发方法

3.1 数据来源

本研究所使用的乳腺癌数据来自国家癌症中心（NCC）发布的《中国肿瘤登记年报》^[13-15]，时间范围为 2014–2016 年，分别包括 339、388 和 487 个肿瘤登记地区，覆盖

中国大陆全部 31 个省份。纳入人口分别为 2.88 亿、3.21 亿、3.82 亿，占中国 2014–2016 年年末人口的 21.07%、23.35%、27.6%；女性人口 1.42 亿、1.58 亿、1.87 亿。其中，女性乳腺癌新发病例分别为 59,806 例、67,328 例、79,450 例。

年报中的乳腺癌发病率有两种统计方式：（1）发病率又称为粗发病率，是反映人口发病情况最基本的指标，是指某年该地登记的每 10 万人口癌症新病例数，反映人口发病水平；（2）世标率又称为年龄调整率或标准化率（Age-standardized rate, ASR），即指按照某一标准人口的年龄结构所计算的发病率，来消除人口年龄结构对发病水平的影响。由于粗发病率受人口年龄构成的影响较大，因此本研究选取年报中的乳腺癌世标发病率（单位：1/10 万）作为统计指标。

3.2 分类标准

本研究根据《2019 中国肿瘤登记年报》^[15]的中国东中西部分区方法，对乳腺癌发病率进行描述性统计分析。地区分类标准如下：

（1）东部地区包括：北京、天津、河北、辽宁、上海、江苏、浙江、福建、山东、广东、海南 11 个省（市）；

（2）中部地区包括：山西、吉林、黑龙江、安徽、江西、河南、湖北、湖南 8 个省；

（3）西部地区包括：内蒙古、广西、重庆、四川、贵州、云南、西藏、陕西、甘肃、青海、宁夏、新疆 12 个省（市、自治区）。

3.3 数据处理

分别提取 2017、2018 和 2019 年《中国肿瘤登记年报》收录的 2014、2015 和 2016 年肿瘤登记地区的乳腺癌发病率数据，汇总在.xlsx 表中，作为描述性统计的源数据。然后在 ArcGIS10.8 软件中完成数据的空间可视化，作为空间统计分析的源数据。考虑到各肿瘤登记处的统计地区与中国县级矢量数据属性表的可匹配性，本研究的 2014–2016 年县级矢量数据分别包含 339、386 和 485 个区县。

3.4 空间统计分析

地理现象普遍存在空间依赖性，这意味着地理空间上接近的地区或观测点之间往往存在着某种程度的相似性或相关性。为了量化这种空间依赖性，Moran's I 统计量被广泛用于检验全局空间自相关。该指数用于衡量每个空间观测值与其附近观测值的平均相似程度，其取值范围在-1 到 1 之间^[18]。当 Moran's I 大于 0 时，表示存在正的空间自相关，即相邻地区的观测值趋向于相似，数值越大则相关性越明显；当 Moran's I 小于 0 时，则表示存在负的空间自相关，即相邻地区的观测值呈现相反的趋势；而当 Moran's I 等于 0 时，则表明空间上的观测值分布基本呈随机分布，即不存在空间自相关。通过 Moran's I 统计量的计算，有助于深入理解地理现象的空间特征及其在空间上的分布规律，为进一步的空间分析和决策提供科学依据。

Moran's I 公式如下：

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n \sum_{j=1}^n w_{ij} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (1)$$

式中, x 为观测变量, x_i 和 x_j 分别为第 i 个和第 j 个空间单元的属性值 (例如第 i 个县级市的乳腺癌发病率, 第 j 个县级市的乳腺癌发病率), w_{ij} 是变量 i 和 j 之间的空间权重, n 为变量总数。

4 数据结果

4.1 中国乳腺癌发病率描述性统计分析

肿瘤登记是对癌症流行情况、趋势变化和影响因素进行长期、连续、动态的系统性监测, 是制定癌症预防控制策略、开展综合防控研究、评价防控效果的重要基础工作^[15]。2014–2016 年, 国家癌症中心分别纳入 339、388 和 487 个肿瘤登记处, 数量逐年增加, 覆盖面更加广泛, 体现了国家对癌症防控工作和人民生命健康的重视。本研究所使用的区县数量和分布如表 1 所示:

表 2 2014–2016 年肿瘤登记处数量统计表

	2014	2015	2016
数量	339	386	485
增值		47	99

首先对中国县级乳腺癌发病率数据进行描述性统计分析。直方图显示, 在 2014–2016 年间, 中国大多数区县的乳腺癌世标发病率 (ASR) 在 35 以下, 而 ASR 在 15–30 左右的区县数量最多, 表明中国乳腺癌发病率总体处于稳定状态 (图 1)。然而, 通过对箱线图的观察可以发现, 中国东部、中部和西部区县的乳腺癌发病率存在一定差异 (图 2)。

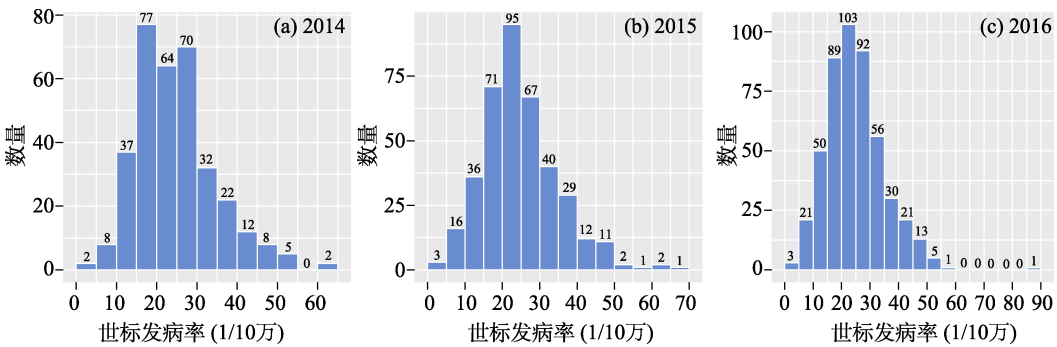


图 1 2014、2015、2016 年中国乳腺癌世标发病率直方图

由表 3 可知, 东部地区的乳腺癌发病率整体高于中部和西部地区, 中值逐年增加, 且方差最大, 表明东部各区县之间发病率存在较大的空间分异现象; 虽然中部地区的发病率中值变化很小, 但方差逐年增加, 提示该地区发病率可能出现了更加明显的空间分异趋势; 西部地区的发病率中值远低于东部和中部地区, 方差基本不变, 说明该地区的发病率一直处于相对稳定的状态。

4.2 中国乳腺癌发病率空间统计分析

2014–2015 年, 中国乳腺癌发病率较高的地区主要分布于辽宁、山东、河南、上海和深圳; 2016 年, 乳腺癌高发区扩大到北京、海南、内蒙和西部的个别地区。而发病率较低

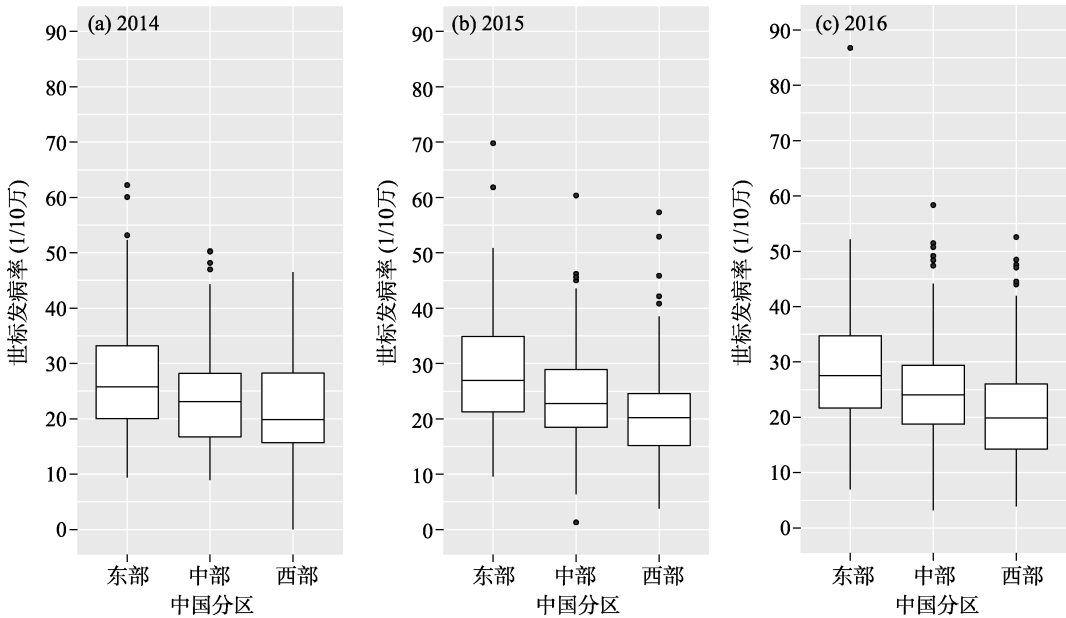


图 2 2014、2015、2016 年中国东中西部乳腺癌发病率箱线图

表 3 2014–2016 年全国东中西部乳腺癌发病率统计表

地区	年份	最大值	最小值	中值	方差
东部	2014	62.23	9.37	25.79	100.57
	2015	69.81	9.57	26.94	101.41
	2016	86.75	6.96	27.54	100.67
中部	2014	50.31	8.91	23.12	82.56
	2015	60.36	1.30	22.75	86.70
	2016	58.35	3.19	24.06	91.90
西部	2014	46.51	0.00	19.84	89.29
	2015	57.33	3.79	20.20	89.28
	2016	52.56	3.88	19.89	89.60

的地区则主要分布于中部和西部地区，以及东部的江苏和福建（图 3）。空间自相关分析结果显示，2014–2016 年，中国乳腺癌发病率 Moran’s I 的 p 值均小于 0.000,1，说明中国乳腺癌发病率始终存在显著空间集聚现象（表 4）。

图 4 表明，相比于 2014 年，2016 年中国大部分地区的乳腺癌发病率都有所增加，特别是京津、辽宁、河北南部、长三角和珠三角，西南和西北的一些地区也出现了发病率上升现象；而河北北部、山东东部和江苏中部等东部地区，安徽中南部、湖南和湖北等中部地区，以及四川、云南、甘肃等西部地区的一些区县，发病率则出现下降。

表 4 2014–2016 年中国乳腺癌发病率全局空间自相关分析统计表

年份	Moran’s I	z-score	p -value	空间模式
2014	0.17	7.12	<0.000,1	clustered
2015	0.12	13.15	<0.000,1	clustered
2016	0.16	20.62	<0.000,1	clustered

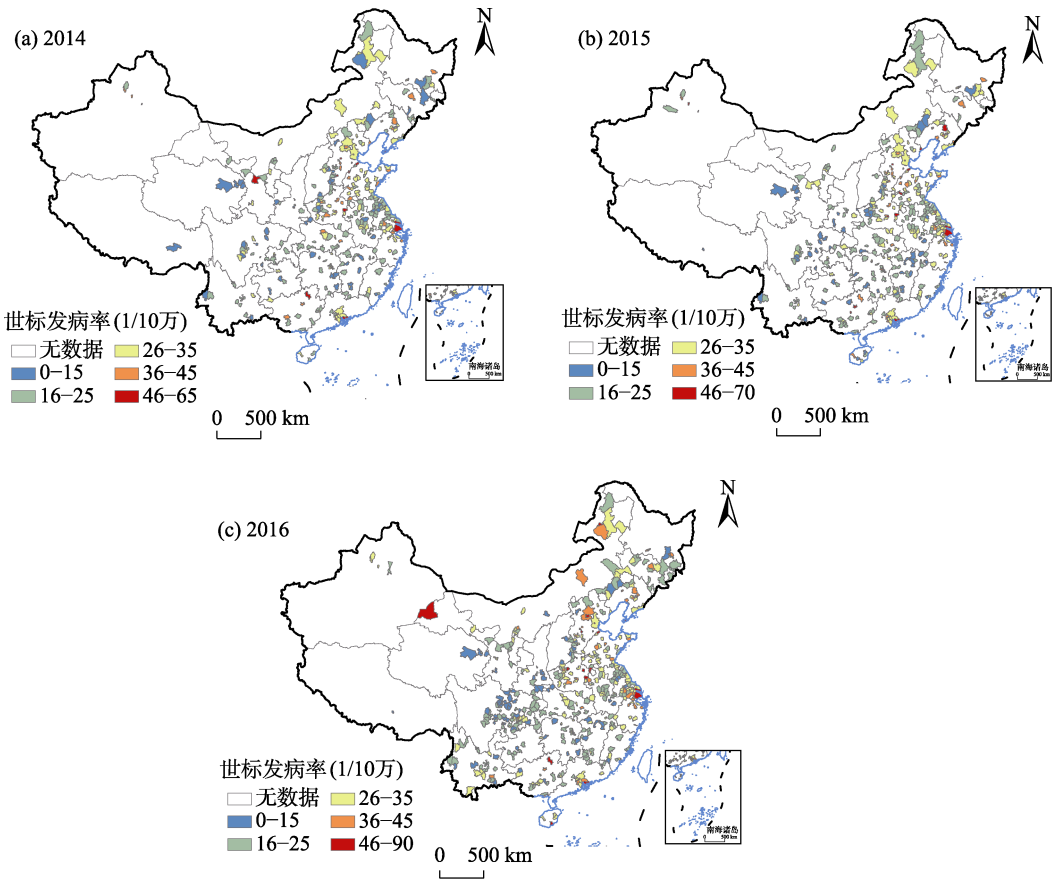


图 3 2014–2016 年中国乳腺癌发病率空间分布图（依据审图号 GS(2020)4619 号的标准地图制作）

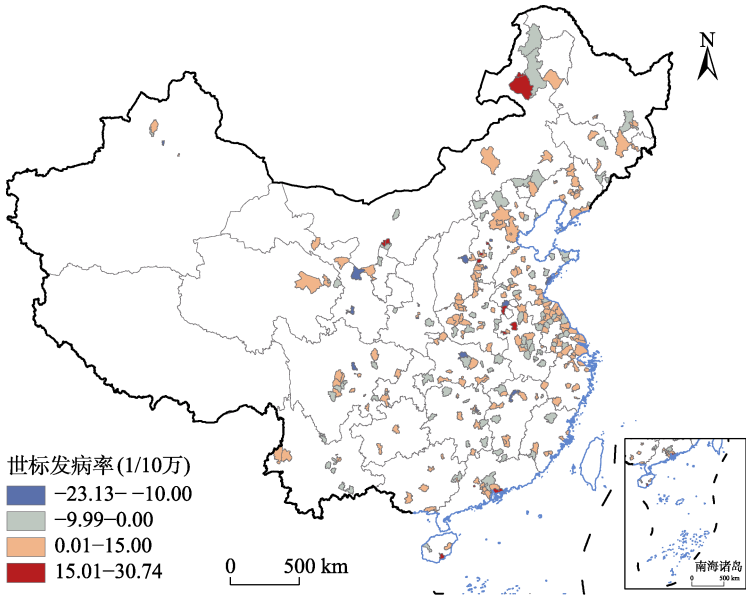


图 4 2014–2016 年中国乳腺癌发病率变化分析图（依据审图号 GS(2020)4619 号的标准地图制作）

5 讨论和总结

本数据集以 2017–2019《中国肿瘤登记年报》为基础，生成 2014–2016 年中国县级乳腺癌发病率矢量数据，提高了空间精确度；以描述性统计和空间统计为分析方法，深入探究了中国乳腺癌发病率的区域差异和空间变化趋势。

结果表明，中国乳腺癌发病率具有明显的区域差异，发病率从高到低依次为东部>中部>西部，存在显著的空间集聚现象。2014–2016 年，中国乳腺癌发病率较高的地区由辽宁、山东、河南、上海和深圳，扩大到北京、海南、内蒙和西部的个别地区；而发病率较低的地区则始终分布于中部和西部地区，以及东部的江苏和福建。2014–2016 年，中国大部分地区的乳腺癌发病率都有所增加，特别是京津、辽宁、河北南部、长三角和珠三角，西南和西北的一些区县也出现了发病率上升现象；而发病率下降的区县主要位于河北北部、山东东部和江苏中部等东部地区，安徽中南部、湖南和湖北等中部地区，以及四川、云南、甘肃和宁夏等西部地区。值得注意的是，部分位于河南、山东等高发区的区县，其乳腺癌发病率出现下降；而河北南部、江苏、安徽北部和宁夏等低发地区的乳腺癌发病率出现了上升趋势。说明在对高发区进行监测、防控与治疗的同时，也要提高对低发区发病率的重视，促进医疗卫生资源的合理配置，增强公众健康意识，逐渐控制和减少由疾病造成的国家和个人负担。

本数据集所覆盖的区县数量约占中国全部区县的 17%，因此可能没有完全反映中国乳腺癌发病率的真实情况，但从整体的数据量来看，本数据集的结果还是有一定代表性的。未来需要在上述结果的基础上，进一步分析乳腺癌发病率的时空分异及其影响机理，从而采取更高效合理的政策与方法来控制和改善由乳腺癌带来的各种负面影响。

作者分工：王若菡对数据集的开发做了总体设计；王若菡采集和处理了数据并撰写数据论文；王若菡、王培涵、徐成东、王伟、王振波做了数据验证。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献

- [1] Sung, H., Ferlay, J., Siegel, R. L., *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [J]. *CA-A Cancer Journal for Clinicians*, 2021, 71(3): 209–249.
- [2] Li, T., Mello-Thoms, C., Brennan, P. C. Descriptive epidemiology of breast cancer in China: incidence, mortality, survival and prevalence [J]. *Breast Cancer Research and Treatment*, 2016, 159(3): 395–406.
- [3] Zheng, R., Zhang, S., Zeng, H., *et al.* Cancer incidence and mortality in China, 2016 [J]. *Journal of the National Cancer Center*, 2022, 2(1): 1–9.
- [4] Lei, S., Zheng, R., Zhang, S., *et al.* Breast cancer incidence and mortality in women in China: temporal trends and projections to 2030 [J]. *Cancer Biology & Medicine*, 2021, 18(3): 900–909.
- [5] Wang, J. F., Zhang, T. L., Fu, B. J. A measure of spatial stratified heterogeneity [J]. *Ecological Indicators*, 2016, 67: 250–256.

- [6] Fei, X., Lou, Z., Christakos, G., *et al.* A geographic analysis about the spatiotemporal pattern of breast cancer in Hangzhou from 2008 to 2012 [J]. *Plos One*, 2016, 11(1): 1–13.
- [7] Song, M. J., Huang, X. X., Wei, X. Q., *et al.* Spatial patterns and the associated factors for breast cancer hospitalization in the rural population of Fujian Province, China [J]. *BMC Womens Health*, 2023, 23(1): 247–255.
- [8] Huo, Q., Zhang, N., Wang, X., *et al.* Effects of ambient particulate matter on human breast cancer: is xenogenesis responsible? [J]. *Plos One*, 2013, 8(10): e76609–e76615.
- [9] Yu, Q., Zhang, L., Hou, K., *et al.* Relationship between air pollutant exposure and gynecologic cancer risk [J]. *International Journal of Environmental Research and Public Health*, 2021, 18(10): 5353–5366.
- [10] He, R., Zhu, B., Liu, J., *et al.* Women’s cancers in China: a spatio-temporal epidemiology analysis [J]. *BMC Womens Health*, 2021, 21(1): 116–129.
- [11] Xia, C., Kahn, C., Wang, J., *et al.* Temporal trends in geographical variation in breast cancer mortality in China, 1973–2005: an analysis of nationwide surveys on cause of death [J]. *International Journal of Environmental Research and Public Health*, 2016, 13(10): 963–978.
- [12] Hu, M. Y., Jiang, C., Meng, R. T., *et al.* Effect of air pollution on the prevalence of breast and cervical cancer in China: a panel data regression analysis [J]. *Environmental Science and Pollution Research*, 2023, 30(34): 82031–82044.
- [13] 国家癌症中心. 2017 中国肿瘤登记年报[M]. 北京: 人民卫生出版社, 2018.
- [14] 国家癌症中心. 2018 中国肿瘤登记年报[M]. 北京: 人民卫生出版社, 2019.
- [15] 国家癌症中心. 2019 中国肿瘤登记年报[M]. 北京: 人民卫生出版社, 2021.
- [16] 王若菡, 王培涵, 徐成东等. 中国乳腺癌发病率空间分布数据集(2014–2016)[J/DB/OL]. 全球变化数据仓储电子杂志, 2024. <https://doi.org/10.3974/geodb.2024.06.06.V1>. <https://cstr.escience.org.cn/CSTR:20146.11.2024.06.06.V1>.
- [17] 全球变化科学研究数据出版系统. 全球变化科学研究数据共享政策 [OL]. <https://doi.org/10.3974/dp.policy.2014.05> (2017 年更新).
- [18] Moran, P. A. P. Notes on continuous stochastic phenomena [J]. *Biometrika*, 1950, 37(1/2): 17–23.