淮河流域上消化道肿瘤与环境污染的模型分析

戚晓鹏[1,2],计 伟[1],任红艳[1],郭 岩[2],周脉耕[2],杨功焕[2],庄大方[1*]
(1. 中国科学院地理科学与资源研究所,北京 100101;
2. 中国疾病预防控制中心,北京 100050)

摘要:自20世纪70年代后期以来,淮河流域不断遭受工业点源污染和其他面源污染,媒体也陆续报道了淮河流域"癌症村"的出现。本文探讨了淮河流域14个监测县5810个行政村的消化道肿瘤与环境因子之间的空间分布规律。作者从流域和行政区划等多维空间角度出发,通过全局的最小二乘法线性回归和稳健回归对环境因子进行筛选分析,以局部地理加权回归方法探测各类环境因子,在不同地区对贝叶斯调整的上消化道肿瘤死亡率的影响程度,建立了消化道肿瘤死亡的风险评估模型,其中,包括地表水水质等级、浅层地下水质量分级、河网密度、土壤多环芳烃含量分级、化肥施用量和经济密度等6类环境危险因素。根据局部回归模型中各监测点环境因子的回归系数和统计学检验结果,提取出当地主要的环境影响因素。从14个监测县区总体上看,地表水水质等级和GDP与肿瘤呈负相关,其他环境因子均与肿瘤死亡存在正相关。但从局部角度看,不同地区环境影响因子种类和影响强度有较大差别。其中淮河流域江苏段以化肥施用量、土壤多环芳烃含量、GDP和河网密度为主要影响因子,安徽段以土壤多环芳烃含量和化肥为主,河南段主要是以地下水质量分级、河网密度和化肥为主,同时河南沈丘县地表水水质等级对当地影响较大。山东段虽然也探测出来部分环境危险因子的存在,但没有发现其与肿瘤死亡的关联关系,尚需进一步深化研究。

关键词:上消化道肿瘤;环境因子;模型分析
DOI:10.3724/SP.J.1047.2012.00432

1 引言

恶性肿瘤是严重威胁人类生存与社会发展的重大疾病,也是21世纪中国和世界最严重的公共卫生问题之一。目前,它已经成为人类死亡构成的重要病因[1],每年全世界约有700多万人死于癌症。根据WHO"世界卫生统计2008"报告,预计全球中、低等收入国家25年后,非传染性疾病的人群死亡率将明显升高;全球癌症死亡率将由2004年的74/10万,上升到118/10万(2030年),仅次于脑血管疾病,死亡率居全死因的第2位。著名肿瘤研究专家、中国工程院院士程书钧认为,世界上新的研究发现,虽然肿瘤的发生是环境因素和人体内遗传因素相互作用的结果,但从总体上来看,环境因素对某些肿瘤的发生更为重要。

淮河流域地处长江流域和黄河流域之间,地跨豫、皖、苏、鲁4省,全长1000多km。随着流域经济快速发展和城市化进度加快,流域水体污染日趋严重,水污染事件时有发生[2]。历史资料显示,1975年淮河发生首次污染,1982年发生第二次污染。进入20世纪90年代,污染事件频繁发生。1992、1994、1995年沙颍河、淮河连续发生大面积水污染事故,对沿淮广大地区工农业生产和城镇供水安全造成严重威胁。随着淮河流域污染的逐年加重,近年来有关癌症高发村的报道时常见诸报端。淮河流域的环境污染与癌症高发的报道出现以来,党中央国务院予以高度重视,温家宝总理、曾培炎副总理等领导对此多次给予重要批示,要求"建立有关部门、沿淮地方政府协作机制,进一步摸清情况,制定综合措施,加强环境治理,加大癌症防治工

收稿日期:2012-05-13;**修回日期**:2012-07-12.
基金项目:"十一五"科技支撑项目"淮河流域水污染与肿瘤的相关性评估研究"(2006BAI19B03)。
作者简介:戚晓鹏(1975-),女,黑龙江省哈尔滨人,博士研究生,副研究员,研究方向为公共卫生信息化以及GIS空间分析在公共卫生领域的应用研究。E-mail:caroline_qi@163.com
*通讯作者:庄大方(1963-),男,博士生导师,研究员,研究方向为资源环境数据库建设与土地利用变化动态监测技术的研究。E-mail:zhuangdf@lreis.ac.cn

作力度”。淮河流域癌症综合防治,是一个复杂的系统工程。本文重点关注历史环境污染物对人群上消化道肿瘤的影响,以探索影响肿瘤发生的相关外部环境因素,将环境因素与人群健康紧密联系在一起,利用传统统计学、GIS 空间统计学方法和生态学研究方法,从多角度多尺度研究环境因素对人群上消化道肿瘤发生造成的影响。

2　研究区数据与环境因素

根据环境监测资料中水质监测断面的位置、结果,以及当地肿瘤患病资料,在淮河流域 4 个省份选择以下 14 个县区(见图 1),共计 5810 个村庄作为研究区,其中包括江苏射阳县、金湖县和盱眙县;安徽灵璧县、埇桥区、颍东区、寿县和蒙城县;山东汶上县和巨野县;河南西平县、扶沟县、沈丘县和罗山县。

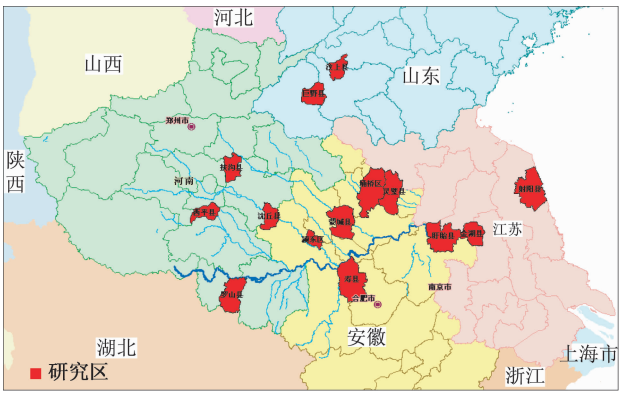

图 1　研究区分布图

Fig. 1　Map of the study areas

2.1　数据处理

本次研究死亡数据来源于 3 部分调查:(1)寿县、灵璧县、罗山县、扶沟县、西平县、盱眙县、射阳县、巨野县、汶上县的死亡资料来源于 2005 - 2006 年死因回顾调查;(2)埇桥区、金湖县、颍东区、蒙城县的死亡资料来源于 2004 - 2005 年全国第三死因回顾抽样调查;(3)沈丘县除 44 个村落为 2002 年 7 月- 2005 年 7 月淮河Ⅰ期的数据外,其他村落数据也均来源于 2004 - 2005 年全国第三死因回顾抽样调查。本论文重点研究死因调查中诊断为上消化道死亡的病例,其中,包括胃癌、食管癌和肝癌 3 种恶性肿瘤。

人口数据来源为:寿县、灵璧县、罗山县、扶沟县、西平县、盱眙县、射阳县、巨野县和汶上县为各县区上报的 2004 - 2006 年分村户籍人口资料。沈丘县、埇桥区、金湖县、颍东区、蒙城县为各县区全国第三死因回顾抽样调查时上报的 2003 - 2005 年分年龄、性别户籍人口资料。

由于在村层面上进行分析时,各行政村人口存在较大差别,例如,在整个研究区域共 5810 个村,其中年平均人口数最少的仅为 30 人,最多的 3 万人,从而导致死亡率的不稳定性。尤其对于小范围人口,如果有一例死亡,就会造成非常高的死亡率。所以本文对村粗死亡率进行空间经验贝叶斯平滑处理[3],按照邻域达到 3 万人其死亡率则相对稳定的假设,综合考虑邻域范围[4-5]、邻域人口、邻域死亡率、邻域死亡率均数和方差,构建风险估计模型。最后将贝叶斯平滑处理后的村死亡率(见图 2)作为下一步建模的肿瘤数据源。

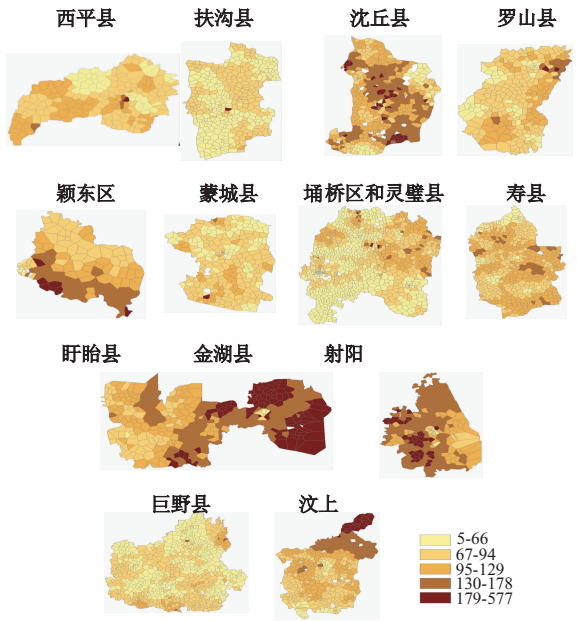

图 2　上消化道肿瘤分村局部贝叶斯调整死亡率(1∶10 万)分布图

Fig. 2　Map of Bayes adjusted mortality of upper digestive tract cancer at village level (1∶100 000 scale)

2.2　环境因子的遴选

从淮河流域的污染特点看[6],可以大体总结为工业排放的点源污染造成地表水、地下水和土壤污染;大量耕地面积,使农药化肥的使用量严重超标,增加了其有害成分在食品和环境的蓄积,导致面源

污染严重;人口密度增加导致生活污水、生活垃圾污染持续升高;经济欠发达导致的粗放式生产增加,环境压力加大等。基于上述淮河流域污染的背景研究,筛选出具有代表性的因素。本次环境污染的影响因素主要包括:地表水水质等级、浅层地下水质量分级、河网密度、土壤多环芳烃含量等级、农药施用量、化肥施用量、经济密度和人口密度等(见图 3)。

3 上消化道肿瘤的环境影响因子分析

3.1 肿瘤的环境影响因子

(1)地表水的质量

据 1997 年的常规环境监测资料中的水质监测数据(来源于 1997 年淮河片水资源公报,共 100 个监测点,分布在淮河干流和一级支流),选取与肿瘤发生有密切关系的化学需氧量(CODMN)浓度、5 日生化需氧量(BOD5)浓度、非离子氨浓度和挥发酚浓度 4 个监测指标。根据 GB3838 标准,以这 4 项指标中最差的一项为准,重新设定监测点水体等级,获得等级为 2 - 6 级(代表 II 类、III 类、IV 类、V 类和劣 V 类)。为了获取河流污染对周边人群健康影响的估计,需要根据样点进行表面的近似模拟,生成覆盖整个淮河流域的地表水水质影响等级图。不规则三角网(Triangulated irregular networks,TIN)插值原理是彼此相邻不重叠的三角形组成的表面,通过在一个三角表面对高程数据进行简单或多项式插值,估计任何位置的表面值。与栅格插值相比,TIN 更好地保存表面要素的精确位置和形状。为了体现距离河道越近,其受河流水质影响的机会越大,本研究采用 TIN 方法对河流的水质等级进行插值。

(2)浅层地下水的质量

2000 年至 2003 年期间,由国土资源部组织开展的新一轮全国地下水资源评价工作。新一轮的全国地下水资源评价主要对象是浅层地下淡水(含水层底板埋深小于 50m),不仅评价了地下水资源的数量,同时也评价了地下水资源的质量。本文所使用的淮河流域地下水质量分布图即来源于此次调查产生的图集[7]。按照国家地下水质量分级标准,将其分为 4 级:第一级为可供引用的地下水,第二级是适当处理后可供饮用的地下水,第三级是不宜直接饮用,但可供工农业利用的地下水,第四级是不宜直接利用的地下水。

(3)河网的密度

淮河流域水体污染比较严重,20 世纪 90 年代地表水监测点主要集中在干流和主要支流上,大部分都处在 V 类和劣 5 类污染状况,但监测点数量较少,而且小支流水质监测缺失,很难勾画出淮河流域地表水污染的总体分布状况。考虑到淮河流域 90 年代污染重、面积广,大部分小支流和沟渠也可能受到较大影响等特点,希望借助河网密度来弥补这一缺陷。通过 TM3 遥感影像和地形提取的河网图层[8],计算每平方公里水体的总长度,作为河网密度,表示为 1km 栅格像元邻域内的线状要素的密度。密度的计量单位为:长度单位(km)/面积单位(km^2)。使用搜索半径以各个栅格像元中心为圆心绘制一个圆。对每条线上落入该圆内的部分长度进行求和,然后将所得的总和除以圆面积。

(4)土壤的多环芳烃含量

据 1997 年中国环境状况公报:“我国耕地污染较重,有 1000 万 hm^2 耕地受到不同程度的污染[9],这占当年 13623 万 hm^2 耕地面积的 7.34%。土壤污染具有隐蔽性和滞后性等特点,它是历史环境污染在土壤中逐年累积的结果,所以,能够反映多年的污染状况。土壤污染包括重金属污染、农药和有机物污染等多种类型。其中,多环芳烃(PAHs),基本为致癌物,所以本文利用淮河流域 21 世纪初土壤多环芳烃抽样调查等级分布指标,级别越高,其 PAHs 的含量越高。

(5)农药和化肥的施用量

农药、化肥可以通过人的皮肤、呼吸道和消化道进入人体,但主要途径是通过人们直接食用含有未完全降解的农药、化肥残留物的农产品,以及农业生产者直接接触。农药和化肥的使用主要集中在耕地,根据单位耕地上的化肥和农药使用率一致的基本假设,利用 1996 年化肥、农药施用量淮河流域分县数据集,以及通过我国土地利用数据[10]提取的耕地类型图层数据,进行空间化处理,分别得到淮河流域的农药和化肥施用量(t)的 1km 栅格数据。

(6)经济密度和人口密度

1995 年普林斯顿大学经济学家 Grossman 等通过实证研究发现,环境污染程度与国民经济收入之间呈倒 U 型曲线关系,即在经济发展的初期,GDP 的提高会导致环境恶化;当经济进一步增长,收入水平提高到一定水平后,收入将伴随环境的改

善而提高。本研究认为在 1978 年改革开放以后，各地经济开始复苏，并逐步进入到快速发展阶段，在 20 世纪 90 年代，一些发达地区逐步关注经济发展造成的环境问题，开始向可持续发展方向转变，由粗放式生产方式向集约型生产方式转变。所以，本研究引入 2000 年每平方公里 GDP 产值这个指标作为经济密度，对间接地分析环境污染有非常重要的意义。人口密度是单位面积土地上居住的人口数[11]。它是表示世界各地人口的密集程度的指标。通常以每平方公里内的常住人口为计算单位。本文采用 2000 年人口密度指标，随着人口密度增加，将导致生活污染增加，也是反映环境污染的间接指标。

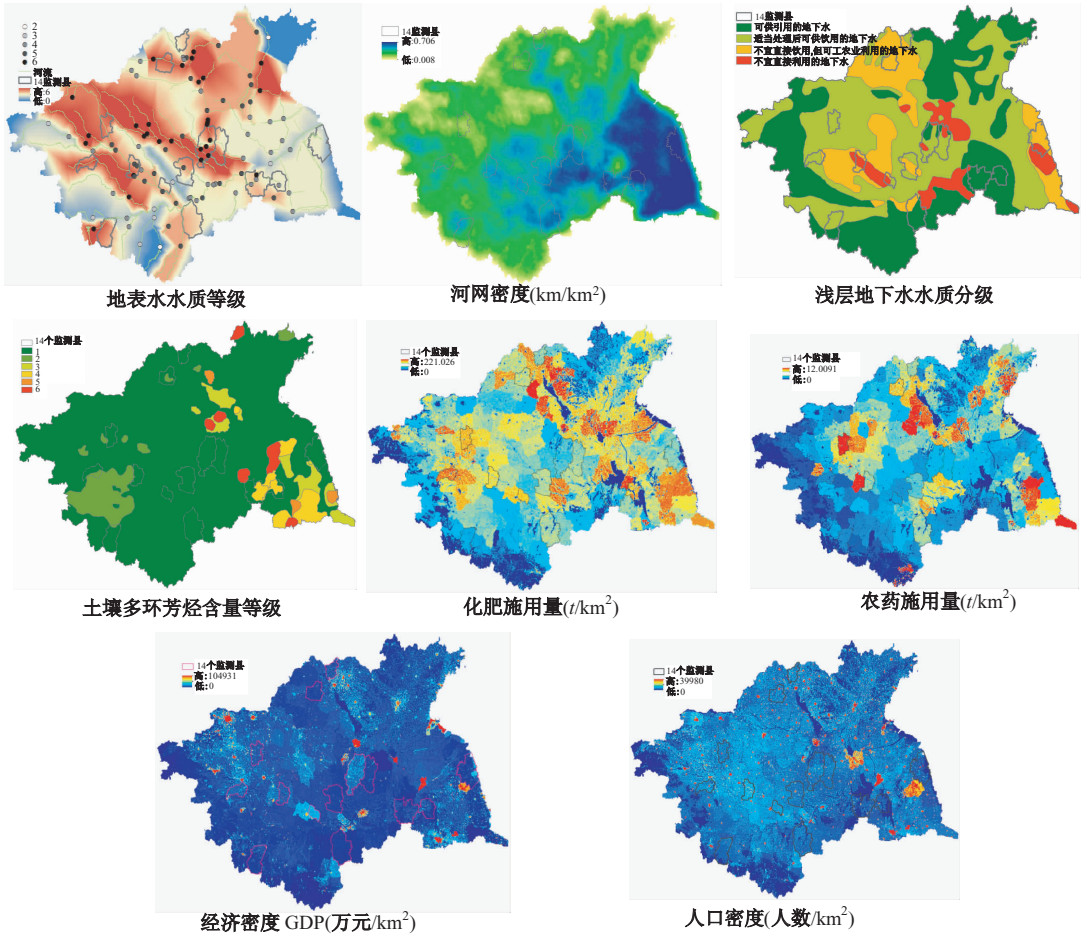

图 3　上消化道肿瘤相关环境影响因子分布图

Fig. 3　Map of distribution of suspected environmental impact factors of upper digestive tract cancer

3.2　分析模型与应用

生态学研究（Ecological study）是在群体的水平上研究某种因素与疾病之间的关系[12-13]。空间分析方法在流行病的关联分析中应用较广，如王劲峰等[14]应用地理探测器研究和顺县神经管畸形与相关环境危险因素的关系。其中，包括 4 种探测器：风险探测器（用于探测环境风险在哪），因素探测器（识别造成环境风险的因素），生态探测器（揭示相关关系），交互探测器（揭示因子之间是否存在交互作用）。本文采用传统统计学与空间统计学相结合的方法，充分发挥两者的优势。

为了研究环境因子与上消化道肿瘤之间的关系，首先，采用普通最小二乘法回归（Ordinary Least Squares，OLS），对解释变量进行筛选。OLS 含义是通过让回归方程计算值（或称估计值）和实际值间差值的平方和最小来建立回归方程的方法[15]。其次，根据自变量筛选过程中获取的诊断信息对模型进一步校正。如两个因素的强关联性可能会造成回归模型的共线性增强；针对异常值所带来的残差非正态分布，通过稳健回归进行校正。该方法的主要目的是检测异常点，并在有异常点的情况下给出模型的稳健估计。其基本思想是对不同数据

点给予不同权重，残差小的点给予较大权重，残差大的点给予较小权重，以减少异常值对模型的影响[16]。

对自变量表现出的空间非平稳性，即回归参数在不同地理位置上表现出差异性[16]。如各类污染物在不同地区对人群健康的影响程度是变化的。所以，这种回归参数是随地理位置变化的，如果仍然采用全局空间回归模型，得到的回归参数估计将是回归参数在整个研究区域内的平均值，不能反映回归参数的真实空间特征。本文在 OLS 和稳健回归的基础上，引入地理加权回归模型（Geographically Weighted Regression，GWR），它是对普通线性回归模型的扩展[17]，将数据的地理位置嵌入到回归参数之中，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 β_0，而是利用邻近观测值子样本数据信息进行局域回归估计而得，随着空间局部地理位置 i 变化而变化的 β_i，GWR 如式(1)表示：

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{k} \beta_j(u_i, v_i) x_{ij} + \varepsilon_i \quad (1)$$

其中，(u_i, v_i) 为第 i 个采样点的地理坐标，β_j 是第 i 个采样点的第 j 个回归参数，是地理位置的函数。GWR 可以对每个观测值估计出 k 个参数向量的估计值，ε_i 是第 i 个区域的随机误差。GWR 模型的核心是空间权重矩阵[18]，它是通过选取不同的空间权函数来表达对数据间空间关系的不同认识。本文采用最为常见的截尾型 bi-square 函数来计算权重，公式如式(2)：

$$w_{ij} = \exp(-(d_{ij}/b)^2) \quad (2)$$

式中，b 是描述权重与距离之间函数关系的非负衰减参数，称为带宽（Bandwidth），带宽越大，权重随距离增加衰减的越慢，带宽越小，权重随距离增加衰减的越快[19]。最优带宽的选取一般有 3 种方法，第一种是交叉验证方法（Cross-Validation，CV），第二种是 Akaike 的信息准则法（Akaike information criterion，AIC），第三种是贝叶斯信息准则（Bayesian Information Criterion，BIC）。本文选择第二种 AIC 方法，即依据极大似然原理，似然函数越大的估计量越好，因为 AIC 是似然函数的对数乘以－2 再加上惩罚因子 2q，因而选择使 AIC 达到最小的模型是“最优”模型。由于本文 14 个监测县 5810 个村点，在研究区域内分布疏密不均，则上述方法求出的最优带宽可能会出现有些回归点周围的数据点过少，导致回归参数估计方差太大，精度很低，甚至出现观测值小于未知参数的现象。为了避免这种情况出现，我们在数据密集地方采用较小的阈值或带宽，而在数据稀疏的地方采用较大的阈值或带宽。最后对模型进行检验，回归方程空间非平稳性的 AIC 比较，就是将具有相同自变量和因变量形式的地理加权回归模型和普通线性回归模型，分别根据其残差平方和与有效参数个数，计算 AIC 值。若 $AIC_{GWR} - AIC_{OLS} > 3$，则拒绝 H_0，说明存在空间不稳定性，GWR 模型比 OLS 模型更接近真实模型。

4 环境因素与肿瘤关联模式结果分析

4.1 OLS 和稳健回归因变量筛选结果

根据容忍度（TOL）越小，膨胀因子（VIF）越大，发生共线性的可能性也越大的原则，去除农药变量。人口密度和单位面积的 GDP 成明显的左偏态分布，与肿瘤的相关性不明显。经 log 转换后，其分布有明显改进，近似正态分布。

将局部贝叶斯调整的上消化道肿瘤死亡率作为因变量 Y，除农药外其他 7 类环境因素作为自变量 X，通过逐步 OLS 回归，计算相关统计量（详见表 1），并进行 OLS 模型诊断（详见表 2、表 3）。

表 1 淮河流域上消化道肿瘤与环境因素 OLS 回归模型分析结果

Tab. 1 OLS regression model analysis results of upper digestive tract cancer and environmental factors in Huaihe River watershed

变量	自由度	参数估计	标准误差	t 值	Pr > |t|	稳健标准误差	稳健 t 值	稳健 Pr	膨胀因子 VIF[1]
截距	1	9.58	5.11	1.87	0.06	5.47	1.75	0.08	—
人口	1	−1.33	0.74	−1.78	0.07	0.77	−1.72	0.08	1.38
GDP	1	0.18	0.59	0.30	0.76	0.60	0.29	0.77	1.34
地表水	1	−1.16	0.05	24.74	0.000*	0.04	26.16	0.000*	1.08
地下水	1	4.52	0.52	8.74	0.000*	0.57	7.88	0.000*	1.16
土壤	1	12.95	0.71	18.16	0.000*	0.91	14.16	0.000*	1.12
化肥	1	0.50	0.04	12.17	0.000*	0.04	11.29	0.000*	1.04
河网密度	1	90.49	4.22	21.46	0.000*	4.94	18.33	0.000*	1.09

表 2　淮河流域上消化道肿瘤与环境因素 OLS 模型诊断

Tab. 2　OLS model diagnosis of upper digestive tract cancer and environmental factors in Huaihe River watershed

输入要素	smooth16	因变量	LOCAL_R
观测值个数	5810	Akaike 的信息标准(AICc) [2]	57684.63829
R 平方的倍数[2]	0.245781	校正 R 平方[2]	0.244871
联合 F 统计量[3]	270.103325	Prob(>F),(7,5802)自由度	0.000000*
联合卡方统计量[4]	1851.60003	Prob(>卡方),(7)自由度	0.000000*
Koenker (BP)统计量 [5]	33.947719	Prob(>卡方),(7)自由度	0.000018*
Jarque-Bera 统计量 [6]	29097.44639	Prob(>卡方),(2)自由度	0.000000*

表 1、表 2 注:* 在 0.05 水平上具有统计学上的显著性

[1] VIF 较大表明解释变量冗余;[2] 模型拟合度/性能的测量;[3] 显著性水平 P 表示整个模型的显著性水平;[4] 显著性水平 P 表示整个鲁棒模型的显著性;[5] 显著性水平 P 表示标准差有偏离;使用稳健估计;[6] 显著性水平 P 表示残差偏离正态分布

表 3　OLS 模型诊断定义

Tab. 3　Definition of model diagnosis

诊断名称	定义
AIC	Akaike 信息准则: 性能的相对测量,用于比较模型;AIC 越小,表示模型越优
AICc	校正的 Akaike 信息标准: 小样本大小的二阶校正
R^2	R 平方,判决系数: 因变量中可由模型解释的变化比例
AdjR^2	校正 R 平方: 针对模型相对于数据的复杂程度(变量数)而进行校正的 R 平方
F-Stat	联合 F 统计量值: 用于评估整体模型显著性
F-Prob	联合 F 统计量概率(p 值): 所有解释变量均对因变量无影响的概率
Wald	卡方统计量: 用于评估整个鲁棒模型显著性
Wald-Prob	卡方统计量概率(p 值): 此概率使用稳健标准差计算得出,表示所有解释变量均对因变量无影响的概率
K(BP)	Koenker 的标准化 Breusch-Pagan 统计量: 存在异方差(非恒定方差)时,用于测试标准差值的可靠性
K(BP)-Prob	Koenker (BP)统计量概率(p 值): 异方差(非恒定方差)未导致标准差不可靠的概率
JB	Jarque-Bera 统计量: 用于确定残差是否偏离正态分布
JB-Prob	Jarque-Bera 概率(p 值): 残差服从正态分布的概率
$Sigma^2$	Sigma 平方: 误差项的方差的 OLS 估计

从总体上看,模型联合 F 统计量和联合卡方统计量远远大于临界值,均有统计学意义,证明模型整体显著,它对上消化道肿瘤的各环境影响因素的探究具有统计学意义,决定系数为 0.246。通过 Koenker(BP)统计量可以衡量自变量是否存在非平稳性现象。本研究 K(BP)-Prob<0.05,说明方程为非平稳性方程。OLS 对于解释变量计算了概率和稳健概率。非平稳性方程必须采用稳健概率来估计其统计学意义,所以,根据表 1 - 3 的 Robust-Prob<0.05 水准,获得 OLS 结果为除人口密度、GDP 和截距项外,地表水、地下水、水体密度、土壤和化肥等 5 个环境因素都有显著意义。JB-Prob<0.05,说明残差不服从正态分布,并指示模型可能会有偏差。这种偏差可能是因为模型中缺失关键的解释变量,或者是建模数据中存在异常点。如果数据中存在异常值,则影响大的异常值可以使模型化的回归关系背离最佳拟合,从而使回归系数发生偏差。通过分析,得到本研究数据中共存在 135 个异常点,占总样本量的 2.32%。稳健回归结果如表 4。

综合上述分析,最终确定地表水水质等级、浅层地下水质量分级、河网密度、土壤多环芳烃含量分级、化肥施用量和经济密度等 6 类环境因素作为模型的自变量,参与到模型的构建。从全局角度看,环境因素与上消化道肿瘤死亡率的关系如下:地表水水质越差,肿瘤死亡率越低;浅层地下水水质等级越差、土壤多环芳烃含量越高、GDP 越低、化肥施用量越高以及河网密度越高,则肿瘤死亡率越高。

4.2　GWR 分析结果

由于 K(BP)的 Prob 值(概率)小于 0.05, 表示

表 4 淮河流域上消化道肿瘤与环境因素稳健回归模型分析结果

Tab. 4 Robust regression model analysis results of upper digestive tract cancer and environmental factors in Huaihe River watershed

参数	自由度	估计值	标准误	95%可信区间		Chi-Square	Pr>ChiSq
截距	1	7.3459	4.3164	−1.1141	15.806	2.9	0.0888
人口	1	−0.9432	0.6252	−2.1685	0.2821	2.28	0.1314
GDP	1	−1.381	0.4906	−2.3426	−0.4194	7.92	0.0049
地表水	1	−1.2154	0.0391	1.1388	1.292	967.27	<.0001
地下水	1	4.1449	0.4372	3.288	5.0017	89.89	<.0001
土壤	1	13.6917	0.6296	12.4577	14.9256	472.96	<.0001
化肥	1	0.4666	0.035	0.398	0.5352	177.82	<.0001
河网密度	1	100.0248	3.6214	92.927	107.1226	762.9	<.0001

模型具有统计学上的空间非平稳性，既而采用 GWR 进一步比较 6 类环境因子在不同地点对肿瘤的影响程度。全局 OLS 模型与局部 GWR 模型的诊断信息如表 5 所示。与 OLS 相比，GWR 模型的残差平方和更小，误差项更小，调整决定系数增加一倍以上，AIC 减少 255。根据模型稳定性检验，$AIC_{GWR} - AIC_{OLS} = 255 > 3$，所以，拒绝 H_0，认为 GWR 模型比 OLS 模型更接近真实模型。

表 5 OLS 与 GWR 模型诊断比较

Tab. 5 Comparison of OLS and GWR model diagnosis

诊断信息	OLS	GWR
残差平方和	6956515.42	4448064.01
有效参数数量	7.00	32.47
Sigma	34.62	27.75
Akaike 信息标准(AIC)	57684.52	55137.55
决定系数(R^2)	0.25	0.52
调整 R^2	0.24	0.51

6 类环境因子的 GWR 回归系数，以及模型截距在不同的村点有不同的取值。回归系数的大小代表了该因素对当地居民健康的影响程度。通过上述 GWR 模型，对肿瘤和环境因子的局部空间探测，获得不同监测村点的回归方程和各因子的回归系数。针对 4 个省、7 个淮河子流域，以及 14 个监测县，提取出当地主要的环境影响因素，并根据局部回归系数，将环境分为影响较大和影响一般，含义是每增加单位面积的污染物或相关指标，肿瘤增加的幅度根据分位数分为较大和一般(图 4)。红色圆点代表不同县区的主要环境影响因子，深红色为影响较大，浅红色为影响一般。没有列出环境因子

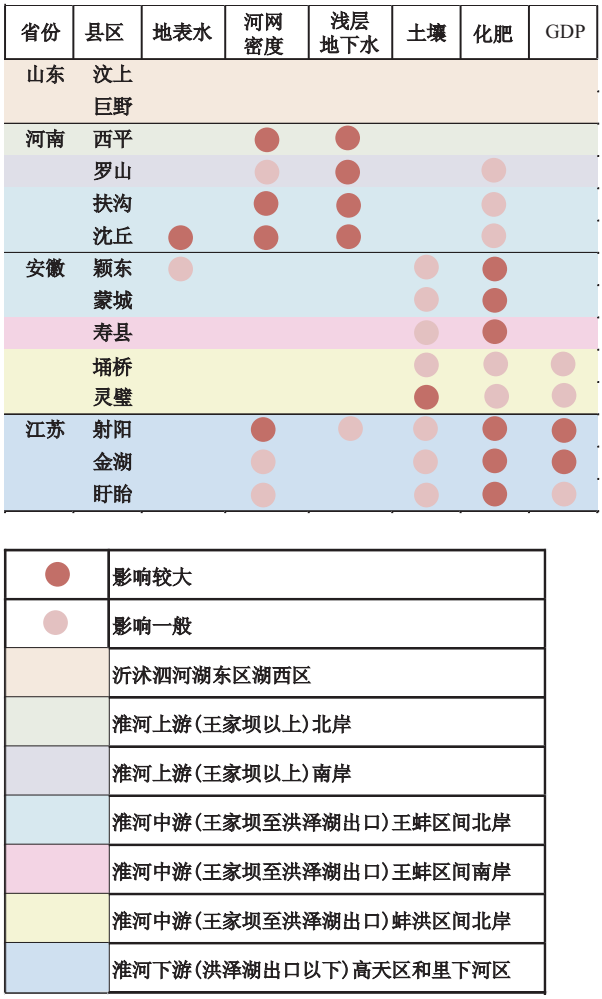

图 4 环境污染相关因素的空间关联性

Fig. 4 Spatial correlation of suspected environmental pollution factors

的县区代表此类因素在当地不显著或对肿瘤发病

没有促进作用。图中不同的背景颜色代表监测县区所在的不同淮河子流域(详见图例)。

从行政区划范围看,山东省包括汶上县和巨野县两个县,地处沂沭泗河流域,在所研究的 6 类环境因素中,没有发现任何对肿瘤发病有正相关作用的因子。河南省包括 4 个研究县区,地处淮河上游和中游上段北岸,水体质量对其影响最大,其中主要以浅层地下水质量和河网密度这两个因素为主,化肥施用量的影响处于一般水平,沈丘地表水水质影响最大。安徽省包括 5 个研究县区,地处淮河中游北岸和中游南岸,相关强度主要以化肥施用量影响为主,其次为土壤的多环芳烃含量,GDP 和地表水水质在个别县区略有影响。江苏省包括 3 个研究县区,地处淮河下游,环境影响因素相对较多。最主要的表现为农药施用量,其次为土壤多环芳烃含量、GDP 和水体密度。地下水水质等级在射阳县呈现一定影响。

从环境因子影响范围上看,化肥影响范围最广,覆盖 3 省 11 个县区。土壤其次,主要集中在淮河中游和下游。河网密度作为间接指示因素主要集中在淮河流域中上游和下游。地下水水质等级对淮河流域中上游影响较大。GDP 所代表的社会经济类型和生产方式对环境的影响主要体现在淮河中下游地区。

从流域角度看[20],淮河下游综合环境因素对肿瘤的影响相对较大,其次为淮河流域中游和上游。沂沭泗河子流域虽然也探测出来部分环境污染因子的存在,但其与肿瘤发生的关系不强,提示这两个县做进一步的调查研究,考虑是否存在肿瘤病人的漏报,同时加入更多的危险因素进行综合分析。

5　结论与讨论

本文在淮河流域 14 个监测县区 5810 个村庄对 8 类环境因子与全人群上消化道肿瘤的空间关联性进行了模型分析:从 14 个监测县区总体上看,地表水水质等级和 GDP 与肿瘤呈负增长,其他环境因子均与肿瘤死亡存在正相关。但从局部角度看,不同地区环境影响因子种类和影响强度有较大差别。一些问题值得进一步研究和讨论。

(1)计算分村的上消化道肿瘤粗死亡率时是采用各地上报的户籍人口。但由于很多地区一部分青壮年常年在外地打工,虽然户籍人口将他们包括在内,但这部分并没有暴露在环境危险因素中。理想情况下应该采用当地常住人口来计算村死亡率。由于在村一级常住人口很难获得,本研究只能用户籍人口进行替代。这种替代会对一部分外出务工较多的县区产生一些影响,从而低估了这些地区的死亡率。

(2)本文筛选出的 8 类环境因子中,只有土壤的多环芳烃含量指标是直接致癌物,而其他都为一些综合质量指标(其中包括一些常规理化检测项,如水质等级)或者一些指代指标(如 GDP 和人口密度),这将大大降低一些因果关系的推断。考虑到历史环境监测中关于直接致癌物数据的匮乏性,所以本文的前提假设是从历年淮河流域污染企业的种类和污染范围看,污染越严重的地方,其排放的直接和间接致癌物的可能性越大。

根据文献显示,肿瘤的发生基本都要经过漫长的过程,10 年、20 年或者更长时间。本文尽量采用一些历史环境数据(采自 1996 - 2003 年的监测数据),肿瘤数据为 2004 - 2006 年。所以,按照暴露—发病的过程,这个时间跨度仍小于 10 年。但从一些污染物,如土壤中污染物,也是常年积累的结果,虽然是 2000 年之后的调查数据,但也能间接反映历史的暴露水平。

地表水水质主要通过淮河干流和一级支流的监测点,以及村庄距河流的距离计算而得。很多污染比较严重的湖泊、水库和沟渠,因为没有获取相关监测资料,所以,即使在其周围有村庄,也要按照距其他大河流的距离来计算。同时水质监测点在淮河流域的东北部分布较少,导致水质指数计算的误差增大。针对本文全局 OLS 回归和稳健回归分析结果中地表水水质越好则肿瘤死亡越高的结果存在两种可能的原因:第一种是地表水水质监测数据精度低,导致其不能全面反映出地表水的污染状况,而我们对于肿瘤的监测数据精度又非常高,达到村级。所以,两者空间尺度的巨大差异将导致我们分析结果的偏差;第二种是地表水对人群健康的影响往往不是通过直接饮用造成的,而是通过污染物在河流的动力作用下在河道底泥、土壤以及食物中的蓄积,间接造成危害。在这种复杂的环境因素综合作用下,会表现出肿瘤与地表水污染的空间不一致性。

假设同一环境污染物,相同剂量,作用在不同

研究县区，其对人群的上消化道肿瘤死亡的发生也会有不同程度的影响，这其中的原因也是相当复杂。首先，人群的易感性会有很大差异，个体的遗传因素也起到很重要的作用；其次，其他危险因素与探测到的这几种环境因素很可能存在交互作用，从而扩大或缩小其对健康的影响结果，如居民的饮食习惯、行为危险因素等[21]；最后，虽然某种环境危险因素存在某一地区，但一部分人群可能没有暴露在这一危险因素中，故没有对健康造成影响，例如，当地下水受到污染，但居民常年都饮用上游的水库水，从而导致其对人群健康影响非常小。所以，本文既用全局 OLS 关联关系的探寻，又采用局部 GWR 进行小范围的环境影响因子与上消化道肿瘤关联关系的探测。从而发现当地肿瘤高发的深层次原因，为预防控制肿瘤的发生制定可行的政策和防治措施。

参考文献：

[1] 陈竺. 全国第三次死因回顾抽样调查报告[M]. 北京：中国协和医科大学出版社，2008.

[2] 李璇. 淮河污染及治污困境分析[J]. 湖北社会科学，2009(5)：3.

[3] Goovaerts P. Geostatistical analysis of disease data：estimation of cancer mortality risk from empirical frequencies using Poisson kriging[J]. International Journal of Health Geographics，2005，4(31)：1－33.

[4] 冯昕，杜世宏，舒红. 空间权重矩阵对空间自相关的影响分析——以我国肾综合征出血热疾病为例[J]. 武汉大学学报(信息科学版)，2011，36(12)：1410－1413.

[5] 刘旭华，王劲峰. 空间权重矩阵的生成方法分析与实验[J]. 地球信息科学，2002，4(2)：7.

[6] 王鑫，梁念，安海蓉. 淮河污染浅析[J]. 中国环境监测，2006，22(1)：3.

[7] 张宗祜，李烈荣. 中国地下水资源与环境图集(M). 北京：中国地图出版社，2004.

[8] 周洪建，史培军，王静爱，等. 近 30 年来深圳河网变化及其生态效应分析[J]. 地理学报，2008，63(9)：969－980.

[9] 国家环境保护总局. 1997 年中国环境状况公报：耕地，2002. http：www. china. com. cn/chinese/zhuanti/hjgb/1007380. htm.

[10] 徐新良，刘纪远，庄大方. 国家尺度土地利用/覆被变化遥感监测方法[J]. 安徽农业科学，2012，40(4)：2365－2369.

[11] Yue T X，Wang Y A，Liu J Y，*et al*. Surface modelling of human population distribution in China[J]. Ecological Modelling，2005，181(4)：461－478.

[12] Glues Forget D，Lebel J. An ecosystem approach to humman health[J]. International Journal of Occupational and Environmental Health，2001，7(2)：1－38.

[13] Tukiendorf A. An ecological analysis of leukemia incidence around the highest 137Cs concentration in Poland[J]. Cancer Causes Control，2001，12(7)：653－659.

[14] Wang J F，Li X H，Christakos G，*et al*. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region，China[J]. International Journal of Geographical Information Science，2010，24(1)：107－127.

[15] 孙振球，徐勇勇. 医学统计学[M]. 北京：人民卫生出版社，2002.

[16] 冯国双，罗凤基. 医学案例统计分析与 SAS 应用[M]. 北京：北京大学医学出版社，2011.

[17] 程娟，万建平. 空间统计在公共卫生事业中的应用[J]. 统计与决策，2009，13(13)：2.

[18] Fotheringham A S，Brunsdon C M C. Geographically weighted regression：a natural evolution of the expansion method for spatial data analysis[J]. Environment and Planning A，1998，30(11)：1905－1927.

[19] 覃文忠. 地理加权回归基本理论与应用研究[D]. 上海：同济大学，2007.

[20] 邢可霞，郭怀成，孙延枫，等. 流域非点源污染模拟研究——以滇池流域为例[J]. 地理研究，2005，24(4)：549－558.

[21] 杨功焕. 中国人群死亡及其危险因素：流行水平、趋势和分布的主要发现[J]. 医学与哲学，人文社会医学版，2007，28(6)：1－5.

Model Analysis of Upper Digestive Tract Cancer and Environmental Pollution in Huaihe River Watershed

QI Xiaopeng[1,2], JI Wei[1], REN Hongyan[1], GUO Yan[2], ZHOU Maigeng[2], YANG Gonghuan[2] and ZHUANG Dafang[1]

(1. *Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing* 100101, *China*;
2. *Chinese Center for Disease Control and Prevention, Beijing* 102206, *China*)

Abstract: The Huaihe River watershed has been suffering from industrial point pollution and other non-point pollution since 1970s. The media has reported the emergence of 'cancer village'. Spatial distribution pattern on upper digestive tract cancer and environmental factors was studied in 14 pilot counties including 5810 villages in Huaihe River watershed. From the multiple perspectives such as watershed and jurisdictional areas, Ordinary Least Squares (OLS) and Robust regression model were used for environmental factors selection. Robust regression model can detect abnormal value and put the weight for each of them. Surface water, phreatic water, river density, soil PAHs, fertilizer, population density and economical density (GDP) were imported into the model. Geographically Weighted Regression (GWR) model was used to locally detect the impact of different environmental factors on Empirical Bayes smoothed upper digestive tract cancer mortality. Through the correlative analysis of them, the risk assessment model of upper digestive tract cancer mortality was developed. Based on the regression coefficient of each environmental factor and statistical test in each pilot, the local main environmental factors were extracted. Even six factors were imported into the global regression model, the impact of each factor was different in distinct areas. In general, all the selected environmental factors show the positive correlation with the upper digestive tract cancer mortality except the surface water quality level and GDP. However the pollution type was diverse in different area based on the regression coefficient of each factor. The main findings were listed as follow: fertilizer amount, soil PAHs, GDP and river density are the main factors in Jiangsu segment in Huaihe River watershed; main factors in Anhui segment included soil PAHs and fertilizer amount; main factors in Henan segment included phreatic water, river density and fertilizer amount, With Shenqiu County, as one of the pilots in Henan, showed the strong positive correlation between surface water and cancer mortality; some kind of environmental risk factors were also detected in Shandong segment, but the result showed no correlation between these risk factors and cancer mortality, which needed to be studied any further.

Key words: upper digestive tract cancer; environmental factors; model analysis