·基础理论与方法·

趋势面分析法在肺癌死亡率地理分布研究中的应用

王晓燕　沈毅　陈坤　缪凡

【摘要】 **目的**　应用趋势面分析法为肺癌死亡率的空间分布规律提供一些线索。**方法**　简述趋势面分析法的原理、具体的操作步骤，利用中国第二次全国死亡调查资料及全国各个省市(区)的地理位置，建立不同性别肺癌死亡率的趋势面函数，选择合适阶次的趋势面函数并做趋势等值线图，进行残差分析并做残差分析图。**结果**　建立了男性肺癌死亡率的五阶趋势面函数(拟合优度为76.06%)以及女性三阶趋势面函数(拟合优度为89.77%)，残差分析提示上海、湖北和天津存在男性肺癌死亡相对增高的因素；浙江、安徽、河南、北京和广西存在男性肺癌死亡相对降低的因素。天津、湖北和广东存在女性肺癌死亡相对增高的因素；浙江、山西、贵州、海南、河南、吉林和内蒙古存在女性肺癌死亡相对降低的因素。**结论**　不同性别的肺癌死亡率空间分布规律是不同的，男性肺癌死亡率东北地区最高，女性肺癌死亡率东南地区最高，其中某些异常点有待于进一步的研究，将对肺癌危险因素的研究提供一些线索。

【关键词】 肺癌；趋势面分析；死亡率

Application of the trend-surface analysis for studing the geographical distribution of lung cancer mortality

WANG Xiao-yan, SHEN Yi, CHEN Kun, MIAO Fan. Institute of Epidemiology and Health Statistics, School of Public Health, Zhejiang University, Hangzhou 310058, China

Corresponding author: *SHEN Yi*, Email: shenyi@zju.edu.cn

【**Abstract**】 **Objective**　To apply trend-surface analysis on lung cancer mortality in China. **Methods**　To overview the theory and approach of trend-surface analysis and to fit the first to fifth order regression equations, where dependent variables were latitude and longitude, and the independent one was the standardized mortality ratio (SMR) of lung cancer for various provinces. Data was from the second mortality survey in the whole country, then proper regression equation was chosen. **Results**　Fifth-order regression equation was established for mortality of lung cancer in males with its fit goodness as 76.06%. The third-order regression equation was established for mortality of lung cancer in females with its fit goodness as 89.77%. There were positive residuals in Shanghai, Hubei and Tianjin, while the negative residuals were in Zhejiang, Anhui, Henan, Beijing and Guangxi for males. There were positive residuals in Tianjin, Hubei and Guangdong, while negative residuals appeared in Zhejiang, Shanxi, Guizhou, Hainan, Henan, Jilin and Neimenggu for females. **Conclusion**　The geographic distribution trend of lung cancer mortality between males and females appeared to be different. The mortality of lung cancer among males was high in the northeast provinces while the mortality of lung cancer for females was high in the southeast areas. In those areas where the residual values were anomalous, further studies need to be carried out.

【**Key words**】 Lung cancer; Trend-surface analysis; Mortality

目前较多研究表明肿瘤的发病率/死亡率存在着空间趋势性[1-4]。趋势面分析是以多元回归分析理论为基础的一种统计方法，能够较准确地反映疾病总的空间分布规律。随着社会的发展、生活环境改变以及人口老龄化等原因，疾病死因谱已发生了很大的改变，肺癌已成为我国居民死亡的主要原因。根据全国两次全死因 3 年回顾调查(1973－1975 年及1990－1992 年)资料显示：肺癌死亡率自 20 世纪 70 年代的7.17/10 万陡增至 90 年代的15.19/10 万，其上升幅度(111.85%)在所有恶性肿瘤中最高，其中男性肺癌死亡率从9.94/10 万升至21.96/10 万，女性肺癌死亡率从4.59/10 万升至8.74/10 万[5]。肺癌的生存期短，我国的肺癌患者 5 年生存率为 10%，西方发达国家的 5 年生存率也仅为 14%，因

作者单位：310058 杭州，浙江大学公共卫生学院流行病学与卫生统计学教研室

通讯作者：沈毅，Email：shenyi@zju.edu.cn

此肺癌死亡率可以反映肺癌的发病情况[6]。本研究旨在通过趋势面分析，揭示我国不同性别肺癌死亡率的空间分布规律，为研究肺癌的病因提供一些线索。

基本原理

设 $Z_j(x_j, y_j)$ 表示所分析现象的特征值，即观测值，其中 (x_j, y_j) 为研究区域内各调查点的坐标。趋势面分析就是把观测值 $Z_j(x_j, y_j)$ 的变化分解成两个部分，即：$Z_j(x_j, y_j) = f(x_j, y_j) + \sigma_j$，式中 $f(x_j, y_j)$ 为趋势值，表示由大区域因素决定的部分，反映了在较大区域范围内 Z 随着 x 和 y 变化的特点；而 σ_j 为剩余值，表示由局部区域因素和随机因素决定的部分，反映了在局部区域范围内 Z 有异于一般规律变化的情况和随机性的干扰所造成的偏差。

趋势面分析法的步骤：

1. 用回归方法求得趋势值和剩余值，即根据已知数据 $Z_j(x_j, y_j)$ 的一个回归方程 $f(x, y)$，使得 $Q = \sum_{j=1}^{n}[Z_j - f(x_j, y_j)]^2$ 达到极小。这实际上是在最小二乘法意义下的曲面拟合问题，即根据观测值 $Z_j(x_j, y_j)$ 用回归分析方法求得一个回归曲面 $\hat{Z} = f(x, y)$。而以对应于回归曲面的值 $\hat{Z}_j = f(x_j, y_j)$ 作为趋势值，以实测值和趋势值之差，即残差 $\hat{Z} - \hat{Z}_j$ 作为剩余值。

2. 建立趋势面回归方程。在趋势面分析中，通常选择多项式作为回归方程。数学表达式：

一阶趋势面函数：$Z_1 = b_0 + b_1 x + b_2 y$

二阶趋势面函数：$Z_2 = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 xy + b_5 y^2$

三阶趋势面函数：$Z_3 = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 xy + b_5 y^2 + b_6 x^3 + b_7 x^2 y + b_8 xy^2 + b_9 y^3$

K 阶趋势面函数：$Z_k = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 xy + b_5 y^2 + \cdots + b_p x^k$

其中 $p = 1/2(k+1)(k+2) - 1$

3. 选择适当阶次的趋势面回归方程。趋势面回归方程对观察值拟合的“好”和“不好”，取决于回归平方和在总离均差平方和中所占的比重，其比重越大表示拟合效果就越好。设 R^2 表示趋势面拟和优度，则 R^2 = 回归平方和/总离均差平方和 100%，R^2 越接近 100% 表示趋势面拟合效果越好。但是趋势面方程不是阶次越高越好，因为随着拟合次数的增加，其通用性和预测性也就越低，计算也越复杂。一般来说，应该根据趋势面方程的显著性检验结果、拟合度和标准误的大小等，结合实际情况综合考虑，选择适当阶次的趋势面方程。本研究采用 SAS 9.0 软件包建立趋势面方程。

4. 绘制等值线图。等值线图的原理：通过趋势面函数公式计算各个观测点 (x_j, y_j) 上 Z 的趋势值 $\hat{Z}_j = f(x_j, y_j)$，并以一定的取值间隔做 Z 的趋势等值线图。等值线相当于 $x - y$ 平面的平行面与趋势面的交线，平行面的高就是 Z 的趋势值，这样立体的趋势面就能以它的等值线图放映到平面上。等值线图采用 SAS 9.0 软件包的 GCONTOUR 过程绘制，然后在 Photoshop 7.0 软件中以中国行政地图为底板叠加而成。

5. 确定正负残差值界限，做残差分析图，进行残差分析[7]。残差指死亡率的实测值与趋势值之差，反映了扣除地理因素作用后局部综合因素（经济、卫生、文化等社会因素和人口特征等）作用下死亡率的变化情况；残差与其标准误之比称学生化残差 (SR_i)，当残差围绕零随机分布，学生化残差在 ±2 之间，认为该方程拟合效果好，无异常点。以实际观察值为横坐标，以残差值为纵坐标，做散点图，如果各点是随机分布的，认为该方程拟合效果好。由残差计算自相关残差区域，正的自相关残差区域提示扣除了地理因素作用后可能存在促使疾病发生的社会因素、人口特征等作用；负的自相关残差区域提示扣除了地理因素作用后可能存在抑制疾病发生的社会因素、人口特征等作用。在残差值确定的条件下可以采用以下的公式：

正残差值上限：$L(+) = \frac{\sum_{j=1}^{m+} e_j(+)}{k}$

负残差值下限：$L(-) = \frac{\sum_{j=1}^{m-} e_j(-)}{k}$

式中 $k = N/x(1, 2, 3, \cdots)$，本研究分析 x 取 2 所得 L 值比较合理。如果各观察点的残差与 L 相比，若正残差小于 $L(+)$ 或者负残差值大于 $L(-)$，认为残差部分的变异主要是随机误差的作用；若正残差大于 $L(+)$ 或者负残差值小于 $L(-)$，认为残差部分的变异主要是局部综合因素的作用。

实例分析

利用我国 20 世纪 90 年代初肺癌死亡率大规模回顾调查资料，数据来源于中国肿瘤防治数据库

(http://cancernet.cicams.ac.cn/)。中国调整死亡率(中调死亡率)采用中国 1964 年人口年龄结构进行调整。各个省市(区)的地理位置(以省会城市所在地为原点),其数据来源于 Google Earth 软件,具体数据见表 1。

1. 趋势面函数的拟合及适度检验:对表 1 资料用 SAS 9.0软件包的 GLM 过程进行趋势面分析,结果见表 2 和表 3。如表 2 所示:男性肺癌死亡率的趋势面函数从第五阶开始,函数的显著性检验才有统计学意义($P<0.05$),拟合度是76.06%,其五阶趋势面函数为:$Z=-15\,077.049\,12+41.048\,53x+2406.558\,81y+0.074\,92x^2-8.984\,94xy-127.305\,51y^2+0.003\,12x^3-0.027\,52x^2y+0.395\,88xy^2+3.316\,68y^3+0.000\,73x^2y^2-0.005\,92xy^3-0.046\,44y^4+0.000\,32y^5$。如表 3 所示:女性肺癌死亡率的一阶到五级阶趋势面函数经显著性检验均有统计学意义。根据拟合优度、拟合优度的增量、变异系数和方程的复杂性等方面的考虑,选择三阶趋势面函数。其方程为:$Z=1703.221\,917-22.159\,705x-66.765\,288y+1.169\,709xy-0.074\,334y^2+0.000\,670x^3-0.005\,261x^2y$。

2. 趋势面函数的等值线图:如图 1 男性肺癌死亡率五阶趋势面所示,我国东北地区男性肺癌死亡率最高,由东北向西南地区过渡肺癌死亡率有下降趋势,肺癌死亡率由南向北有增长趋势,由西向东有下降趋势。如图 2 女性肺癌死亡率三阶趋势面所示,我国东南沿海一带女性肺癌死亡率最高,东南向西北地区过渡肺癌死亡率有下降趋势,肺癌死亡率由南向北有增长趋势,但是由东向西方向变化不是很明显。

表1　20 世纪 90 年代初我国 27 个省市(区)肺癌死亡率(/10 万)与地理位置分布

省市(区)	x 经度(东经)	y 纬度(北纬)	男性				女性			
			中调率	五阶趋势值	残差	学生化残差	中调率	三阶趋势值	残差	学生化残差
北京	116°23′	39°54′	17.52	21.78	−4.26	−1.15	13.23	13.81	−0.58	−0.31
天津	117°11′	39°07′	30.86	22.30	8.56	2.06	19.45	13.53	5.92	3.09
河北	114°29′	38°03′	22.97	20.82	2.15	0.49	11.30	11.15	0.15	0.08
山西	112°33′	37°53′	20.59	18.48	2.11	0.55	8.06	9.99	−1.93	−0.99
内蒙古	111°38′	40°48′	26.55	28.66	−2.11	−0.93	9.82	11.35	−1.53	−1.12
辽宁	123°23′	41°48′	11.67	13.55	−1.88	−0.56	19.03	19.31	−0.28	−0.16
吉林	125°18′	43°52′	10.96	11.59	−0.63	−0.20	20.55	22.10	−1.55	−0.89
黑龙江	126°38′	45°45′	34.15	33.34	0.81	0.74	24.94	24.28	0.66	0.57
上海	121°28′	31°12′	35.04	28.82	6.22	2.03	10.97	9.89	1.08	0.65
江苏	118°46′	32°03′	27.35	27.67	−0.32	−0.08	9.62	8.67	0.95	0.50
浙江	120°09′	30°15′	18.15	23.43	−5.28	−1.28	6.38	8.44	−2.06	−1.12
安徽	117°16′	31°51′	19.76	24.89	−5.13	−1.17	6.62	7.82	−1.20	−0.62
福建	119°19′	26°04′	17.67	18.38	−0.71	−0.53	5.42	6.19	−0.77	−0.67
江西	115°53′	28°40′	13.79	14.32	−0.53	−0.14	5.31	5.58	−0.27	−0.14
山东	116°58′	36°39′	25.71	27.08	−1.37	−0.35	11.15	11.24	−0.09	−0.04
河南	113°38′	34°45′	18.13	23.06	−4.93	−1.16	6.79	8.39	−1.60	−0.82
湖北	114°16′	30°34′	25.03	18.01	7.02	1.64	7.46	6.13	1.33	0.70
湖南	112°58′	28°12′	13.70	12.37	1.33	0.33	4.89	4.97	−0.08	−0.04
广东	113°15′	23°06′	22.89	22.09	0.80	0.53	7.40	3.78	3.62	2.04
广西	108°21′	22°47′	15.55	18.15	−2.60	−1.01	6.38	6.25	0.13	0.08
海南	110°20′	20°02′	28.62	27.97	0.65	1.08	3.41	5.18	−1.77	−1.65
四川	104°04′	30°40′	19.79	20.17	−0.38	−0.14	6.83	7.02	−0.19	−0.12
贵州	106°43′	26°34′	17.76	16.87	0.89	0.29	5.03	6.81	−1.78	−0.96
云南	102°42′	25°03′	28.22	27.68	0.54	0.58	12.32	11.22	1.10	0.98
陕西	108°53′	34°15′	14.55	16.72	−2.17	−0.59	5.98	6.76	−0.78	−0.42
甘肃	103°45′	36°04′	10.79	11.09	−0.30	−0.12	5.71	4.96	0.75	0.52
宁夏	106°13′	38°28′	16.25	14.75	1.50	0.53	6.94	6.17	0.77	0.51

表2　20 世纪 90 年代初我国男性肺癌死亡率趋势面分析参数估计

趋势面	拟合优度(R^2)	变异系数(CV)	均方根	F 值	P 值
一阶	0.046 962	33.572 46	7.013 162	0.59	0.562
二阶	0.069 119	35.470 80	7.409 719	0.31	0.900
三阶	0.182 247	35.909 41	7.501 343	0.50	0.840
四阶	0.566 411	29.648 89	6.193 543	1.52	0.224
五阶	0.760 551	23.798 51	4.971 420	2.72	0.045[a]

注:[a]$P<0.05$

表3　20 世纪 90 年代初我国女性肺癌死亡率趋势面分析参数估计

趋势面	拟合优度(R^2)	R^2 增量	变异系数(CV)	均方根	F 值	P 值
一阶	0.617 820	-	36.422 30	3.520 687	19.40	<0.01
二阶	0.861 524	0.243 704	23.437 78	2.265 566	26.13	<0.01
三阶	0.897 678	0.036 154	21.761 45	2.103 526	19.74	<0.01
四阶	0.921 253	0.023 575	21.646 68	2.092 432	13.65	<0.01
五阶	0.938 063	0.016 810	20.735 88	2.004 392	12.98	<0.01

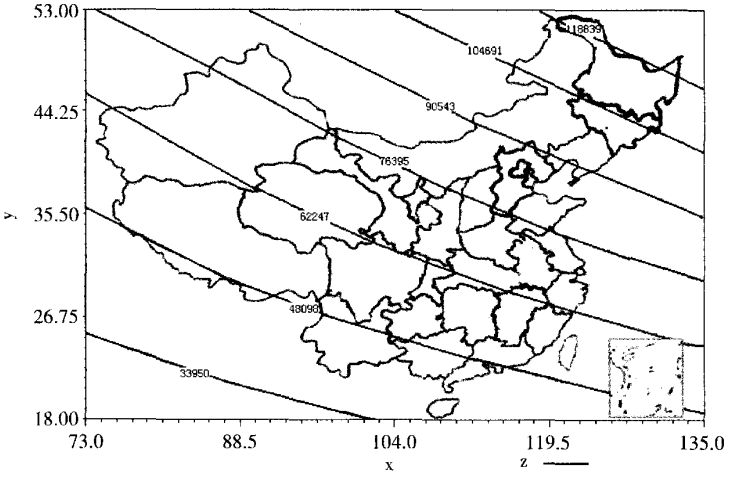

图1 20 世纪 90 年代初我国各省市(区)男性肺癌死亡率(/10 万)的趋势等值线图

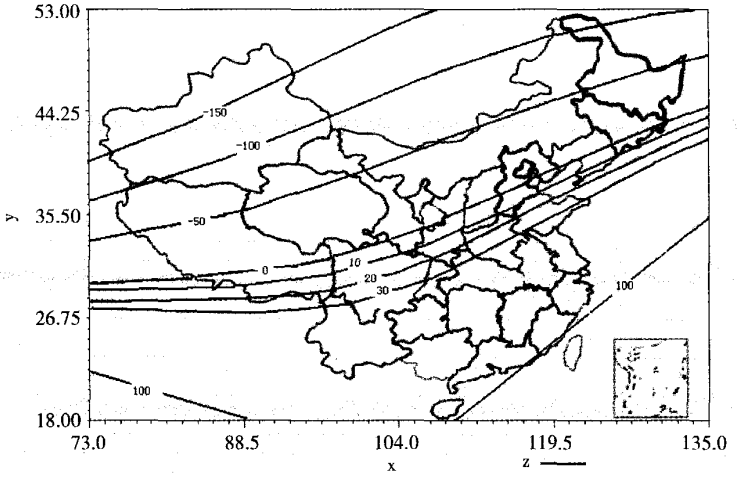

图2 20 世纪 90 年代初我国各省市(区)女性肺癌死亡率(/10 万)的趋势等值线图

3. 残差分析：男性五阶趋势面函数和女性三阶趋势面函数残差分析图分别见图 3 和图 4。其散点分布比较均匀，说明趋势面函数的拟合效果比较好。

将表 1 中 27 个省市(区)的男性残差值划分为正负两组，其中正残差值有12个，正残差值之和为 2.58，负残差值有 15 个，负残差值之和为 -32.60，$k = N/2 = 13.5$，计算所得 $L(+) = 2.41$，$L(-) = -2.41$。12 个正残差值中有 3 个高于 $L(+)$，分别是上海、湖北和天津，15 个负残差值中有 5 个低于 $L(-)$，分别是浙江、安徽、河南、北京和广西。

将表 1 中 27 个省市(区)的女性残差值划分为正负两组，其中正残差值有 11 个，正残差值之和为 16.46，负残差值有 16 个，负残差值之和为 -16.46，$k = N/2 = 13.5$，计算所得 $L(+) = 1:22$，$L(-) = -1.22$。11 个正残差值中有 3 个高于 $L(+)$，分别是天津、湖北和广东，16 个负残差值中有 7 个低于 $L(-)$，分别是浙江、山西、贵州、海南、河南、吉林和内蒙古。

讨 论

趋势面分析是疾病空间分析的主要工具之一，在地理流行病学中得到了较多的应用，但是目前关于肺癌趋势面研究分析在国内未见报道。趋势面分析以多元回归分析理论为基础，从整体出发，分析疾病空间分布规律和局部变异。等值线图为剔除局部和随机变异影响后求得，能较准确地反映疾病地区分布总的变化规律，同时趋势面分析将每一观测值分解成为趋势值和剩余值两部分。趋势值所组成的趋势面表示研究区域的系统性变异，

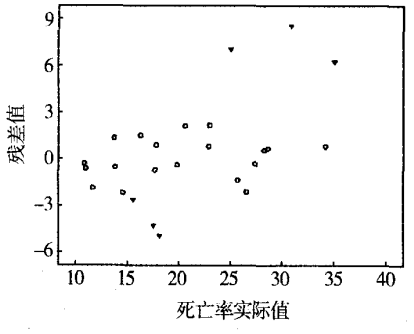

图3 男性肺癌死亡率(/10 万)趋势面分析残差分析

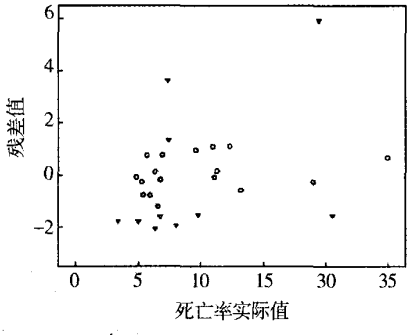

图4 女性肺癌死亡率(/10 万)趋势面分析残差分析

这种变异一般认为是由于环境总的变化或人群的系统属性变化所致。趋势面的剩余值包括随机误差和局部因素所引起的变化，如果趋势面能完全代表观测值，则误差是由随机误差所致。在实际工作中，拟合往往是不完全的，所以误差既包括随机误差，也反映了局部因素的变异。趋势面分析不但能从参差不齐的数据中剔除局部和随机变异的影响，显示大范围区域的系统性变化，研究总的变化规律，而且也能反映出疾病局部地区变异，为病因探讨提供更为重要的线索。特别适用于反映研究区域内较小样本的总体趋势。

曲宸绪等[8]利用反距离加权法以我国 20 世纪 90 年代抽样地区的肿瘤数据预测非抽样地区的死亡率，绘制了全国肿瘤分布地图，较准确地反映了我国肿瘤分布规律。反距离加权法是最常见的空间内插方法之一，属于几何内插法[9]。该方法认为与未采样点的距离最近的若干采样点对未采样点值的贡献最大。但根据该法绘制的肿瘤死亡分布地图不能校正实测数据的波动性。趋势面分析是属于统计内插法[9]，其基本假设是一系列空间数据相互相关，预测值的趋势和周期是与它相关的其他变量的函数。趋势面分析可以剔除局部和随机变异的影响后求得等值线图，较常用的疾病地区分布图能更加准确地放映疾病地区分布总的变异趋势。但是由于数据可获得性的局限，本文未采用各个抽样地区的具体死亡率，而是以各个省份的死亡率做趋势面分析，可能会导致结果不够精确。尽管如此，趋势面分析还是能够反映我国不同性别肺癌死亡率的空间分布规律。

本文应用五阶的趋势面函数拟合男性肺癌死亡率，所建立的方程有统计学意义，拟合优度为 76.06%。等值线图提示男性肺癌死亡率东北地区最高，由东北向西南地区过渡肺癌死亡率有下降趋势，由南向北有增长趋势，由西向东有下降趋势。女性肺癌死亡率的三阶趋势面方程的拟合优度为 89.77%，残差分析图分布比较均匀，亦提示拟合方程效果较好。从等值线图可以看出女性肺癌死亡率以东南沿海一带最高，由东南向西北方向有下降趋势，由南向北有增长趋势，但是由东向西方向的变化不大。

肺癌死亡率在地理分布上有一定特征，我国东北地区及东部沿海较高，而西北和西南地区较低，与周有尚[10]的研究结果相一致，但本文绘制了趋势等值线图，结果更加直观、形象。2002 年 WHO 报告，影响健康的前十大危险因素中，吸烟排在第 4 位[11]。我国 15 岁及以上居民吸烟率为24.0%，男性(50.2%)明显高于女性(2.8%)[12]。本研究显示不同性别的肺癌死亡率的地理趋势不同可能与我国不同性别的吸烟率差异有关，还有待进一步研究。

残差分析结果提示，上海、湖北和天津存在男性肺癌死亡相对增高的因素；浙江、安徽、河南、北京和广西存在男性肺癌死亡相对降低的因素。天津、湖北和广东存在女性肺癌死亡相对增高的因素；浙江、山西、贵州、海南、河南、吉林和内蒙古存在女性肺癌死亡相对降低的因素。这些省份之间的某些因素，如地理环境、生活习惯、经济状况等的差异有待于进一步的研究，可能会对研究肺癌死亡率的危险因素提供一些线索。但是在残差分析中 L 值是人为设置的，随着 x 的不同 L 值会发生变化，影响参差分析的结果。

参 考 文 献

[1] 田承业，王笃宝，薛付忠. 山东省胃癌死亡率地域分布的趋势面分析. 社区医学杂志，2003，1(2)：31-32.

[2] 赵玉婉，陈坤，马新源，等. 大肠癌发病地理特征的趋势面分析. 生物数学学报，2005，20(1)：101-106.

[3] 彭仙娥，史习舜. 应用趋势面分析探索食管癌死亡率的地理分布特征. 海峡预防医学杂志，2003，9(2)：66-67.

[4] 裘炯良，郑剑宁，周健，等. 趋势面分析法在传染病地理分布研究中的应用. 中国热带医学，2004，4(5)：689-691.

[5] 杨玲，李连弟，陈育德，等. 中国肺癌死亡趋势分析及发病死亡的估计与预测. 中国肺癌杂志，2005，8(4)：274-278.

[6] Gillil FD, Samet JM . Lung cancer, 'trends in cancer incidence and mortality'. Cancer Surv , 1994, 19(20): 77-98.

[7] 梁跃武. 残差分析在回归模型诊断中的应用. 中国卫生统计，1991，8(2)：49-51.

[8] 曲宸绪，姜勇，武燕萍，等. 使用反距离权重内插法绘制中国1990年代肿瘤分布地图. 中华流行病学杂志，2006，27(3)：230-233.

[9] 李新，陈国栋，卢玲. 空间内插方法比较. 地球科学进展，2000，15(3)：260-265.

[10] 周有尚. 中国 1990 - 1992 年肺癌死亡流行分布. 中国肿瘤，1997，6(9)：3-7.

[11] WHO. The world health report 2002. Geneva: WHO, 2002: 3-6.

[12] 马冠生，孔灵芝，栾德春，等. 中国居民吸烟行为的现状分析. 中国慢性病预防，2005，13(5)：195-199.

(收稿日期：2006-11-22)

(本文编辑：张林东)