

广州市灰霾致肺癌的因果性分析

刘晓雪,胡三清

(杭州电子科技大学计算机学院,浙江 杭州 310018)

摘要:以广州市为例,研究了灰霾与肺癌的因果性问题,并进一步确定了肺癌的死亡率与灰霾的滞后关系.首先,构建灰霾与肺癌死亡率的线性回归模型;然后,通过计算两个变量之间的格兰杰因果关系值以及新型因果关系值确定两变量是否存在因果关系,再对所求得的因果关系值进行显著性检验;最后,计算不同滞后项的值大小,通过比较不同滞后项所占的比例大小来确定肺癌与灰霾之间的滞后关系.该回归模型的建立与因果关系分析方法的应用证实了灰霾是导致肺癌的死亡率上升的一个重要原因,平均而言肺癌死亡率与灰霾存在 8 年的因果滞后关系.

关键词:新型因果关系;格兰杰因果关系;肺癌;灰霾;大气污染

中图分类号:X-4

文献标识码:A

文章编号:1001-9146(2016)06-0021-04

0 引言

随着经济的快速发展,中国以及世界各国都遭受着严重的环境污染,比如霾污染.肺癌作为全球发病率最高的恶性肿瘤之一,其发病率呈现持续上升的趋势.目前,对于灰霾和肺癌的研究仅仅停留在两者的相关性上.格兰杰因果关系法^[1]是目前最普遍的一种计算因果关系的方法.文献[2]指出了格兰杰因果关系的不足并且提出了新型因果关系的概念.文献[3-4]用大量的实例证实了新型因果关系法比经典的格兰杰因果关系法更能准确地揭示真实的因果关系.本文根据广州市气溶胶光学消光系数(Aerosol Optical Extinction Coefficients, AEC)的历年变化趋势以及广州市的肺癌死亡率变化趋势^[5],建立了气溶胶光学消光系数与肺癌死亡率的线性回归模型,运用因果关系的方法,得出灰霾严重程度是导致肺癌的一个重要外部因果原因.

1 数据分析方法与结果

1.1 实验数据

文献[5]对广州市历年的气溶胶光学消光系数和肺癌死亡率进行了长期的统计.气溶胶光学消光系数是通过在广州市的大气能见度的常规观测中得到的,由于肺癌的发生需要长时间暴露在细小颗粒物上,气溶胶的快速变化水平对疾病没有显著的影响,所以对气溶胶进行 10 年滑动平均的处理.根据文献[5]的研究成果,本文运用 Getdata graph Digitizer 软件^[6],提取得到广州市历年的气溶胶光学消光系数 AEC 的变化趋势图以及历年肺癌的死亡率变化趋势图,如图 1 所示.

从图 1 可以看出,肺癌的死亡率与气溶胶光学消光系数存在一个明显的滞后期.所以肺癌的死亡率与气溶胶光学消光系数之间存在一定的滞后关系.

收稿日期:2016-07-20

基金项目:国家自然科学基金资助项目(61473110);浙江省自然科学基金重点资助项目(LZ13F030002)

作者简介:刘晓雪(1992—),女,浙江温州人,硕士研究生,认知计算与应用.通信作者:胡三清教授, E-mail: sqhu@hdu.edu.cn.

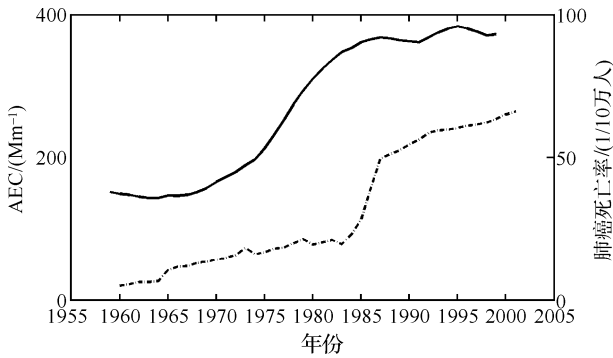


图 1 历年 AEC 变化趋势与肺癌死亡率变化趋势的关系图

1.2 格兰杰因果关系与新型因果关系

格兰杰因果关系可以简要地概括为:考虑 2 个时间序列 $X_{1,t}, X_{2,t}$, 当 $X_{2,t}$ 之前的值提高了预测 $X_{1,t}$ 序列当前值的准确性, 那么 $X_{2,t}$ 对 $X_{1,t}$ 就有因果关系. 自回归模型表示为:

$$\begin{cases} X_{1,t} = \sum_{j=1}^m a_{11,j} X_{1,t-j} + \epsilon_{1,t}, & \text{var}(\epsilon_{1,t}) = \Sigma_1, \\ X_{2,t} = \sum_{j=1}^m a_{22,j} X_{2,t-j} + \epsilon_{2,t}, & \text{var}(\epsilon_{2,t}) = \Gamma_1. \end{cases} \tag{1}$$

其中, $a_{11,j}, a_{22,j}$ 表示自回归模型的系数, m 表示模型的阶数, t 的取值为 $1 \sim N$, N 为样本的容量, $\epsilon_{1,t}, \epsilon_{2,t}$ 为误差项, Σ_1, Γ_1 表示误差项的方差. 此时误差项大小取决于 $X_{1,t}$ 和 $X_{2,t}$ 的自身过去值. 联合回归模型表示为:

$$\begin{cases} X_{1,t} = \sum_{j=1}^m a_{11,j} X_{1,t-j} + \sum_{j=1}^m a_{12,j} X_{2,t-j} + \eta_{1,t}, & \text{var}(\eta_{1,t}) = \Sigma_2, \\ X_{2,t} = \sum_{j=1}^m a_{21,j} X_{1,t-j} + \sum_{j=1}^m a_{22,j} X_{2,t-j} + \eta_{2,t}, & \text{var}(\eta_{2,t}) = \Gamma_2. \end{cases} \tag{2}$$

其中, $a_{11,j}, a_{12,j}, a_{21,j}, a_{22,j}$ 表示联合回归模型的系数, m 表示模型的阶数, t 的取值为 1 到 N , N 为样本容量, $\eta_{1,t}, \eta_{2,t}$ 是预测误差, Σ_2, Γ_2 表示误差项的方差. 格兰杰因果关系表示为:

$$F_{X_2 \rightarrow X_1} = \ln \frac{\Sigma_1}{\Sigma_2}, \tag{3}$$

$$F_{X_1 \rightarrow X_2} = \ln \frac{\Gamma_1}{\Gamma_2}. \tag{4}$$

如果 $\Sigma_2 < \Sigma_1$, 那么 X_2 对 X_1 就有因果关系, 也就是说 X_2 的加入使得对 X_1 的预测更加的准确; 如果 $\Sigma_2 = \Sigma_1$, 那么就说明 X_2 对 X_1 不存在因果关系. 同理可得 X_1 对 X_2 是否存在因果关系. 从格兰杰因果关系定义的式子中可以看出, 格兰杰因果关系只与误差项有关, 而与联合回归模型的系数(部分系数无关, 比如 $a_{11,j}$, 更详细的解析参见文献[2-3]).

文献[2]指出了格兰杰因果关系的不足并且提出了新型因果关系的概念. 根据式(2)对 $X_{1,t}$ 起作用的包括 $\sum_{j=1}^m a_{11,j} X_{1,t-j}, \sum_{j=1}^m a_{12,j} X_{2,t-j}$ 和误差项 $\eta_{1,t}$, 那么 X_2 对 X_1 的新型因果关系表示为:

$$N_{X_2 \rightarrow X_1} = \frac{\sum_{t=m+1}^N (\sum_{j=1}^m a_{12,j} X_{2,t-j})^2}{\sum_{h=1}^2 \sum_{t=m+1}^N (\sum_{j=1}^m a_{1h,j} X_{h,t-j})^2 + \sum_{t=m+1}^N \eta_{1,t}^2}. \tag{5}$$

同理可得 X_1 对 X_2 的新型因果关系:

$$N_{X_1 \rightarrow X_2} = \frac{\sum_{t=m+1}^N \left(\sum_{j=1}^m a_{21,j} X_{1,t-j} \right)^2}{\sum_{h=1}^2 \sum_{t=m+1}^N \left(\sum_{j=1}^m a_{2h,j} X_{h,t-j} \right)^2 + \sum_{t=m+1}^N \eta_{2,t}^2}.$$

(6)

1.3 线性回归模型的阶数选择

常用回归模型的阶数选择准则很多,本文主要应用赤池信息准则(Akaike Information Criterion, AIC)^[7].具体函数表示为:

$$AIC(m) = N \log(\det(\mathbf{\Sigma})) + 2mm^2.$$

(7)

其中, N 表示时间序列的全部样本点数, m 是模型的滞后阶数, n 是构成回归模型的变量的个数, $\mathbf{\Sigma}$ 则表示协方差矩阵. $AIC(m)$ 是一个关于阶数 m 的离散函数,最优阶数应该使得函数值最小. 阶数 m 的选取采取 Schwart 推荐的方法^[8], m 的最大值是 $12(T/100)^{0.25}$ (T 表示样本的容量),计算得 m 的最大取值为 9,然后分别计算 1~9 的 AIC 的值,根据式(7) 计算可得 AIC 值在 m 取为 9 的时候最小,所以构建模型的最佳阶数为 9.

1.4 因果关系统

1.3 节中确定了构建肺癌死亡率与气溶胶光学消光系数的因果关系模型的最佳滞后阶数为 9 阶. 根据格兰杰因果关系和新型因果关系的计算公式,进一步得到了新型因果关系和格兰杰因果关系的值分别为 0.051 5 和 0.934 6.

1.5 显著性检验

为了证明上述因果关系统的有效性,对这 2 个值进行显著性检验.

显著性检验描述为:当计算 X_2 对 X_1 的因果关系统的时候,首先打乱 X_2 序列的顺序,然后计算打乱顺序的 X_2 序列对 X_1 的因果关系统. 重复这个过程 100 次,就得到了 100 个因果关系统的值,然后把得到的 100 个因果关系统值从小到大排序,如果真实的因果关系统值比排序后得到的第 95 个因果关系统值要大,就说明得到的因果关系统是有效的,具有显著性.

根据计算,打乱顺序后的气溶胶光学消光系数与肺癌死亡率序列所计算的新型因果关系统值都比真实的值小,两者计算的格兰杰因果关系统值只有 94 次比真实值小,因此运用格兰杰因果关系统计算的因果关系统值不具有显著性,而新型因果关系统计算的因果关系统值具有显著性. 所以新型因果关系统值具有显著性从而得出灰霾对肺癌有因果影响,而格兰杰因果关系统值不具有显著性从而不能得出灰霾对肺癌有因果影响的结论.

1.6 因果滞后关系的确定

根据前面所述,最佳滞后阶数取为 9,根据新型因果关系计算公式(5),若阶数为 9,新型因果关系统值的分子部分 $\sum_{t=m+1}^N \left(\sum_{j=1}^m a_{12,j} X_{2,t-j} \right)^2$ 有 9 项,要确定 2 个时间序列的滞后关系,把在因果关系统中所占比例最大的项确定为滞后项. 滞后 j 年项的值如下:

$$\text{lagYear}(j) = \sum_{t=10}^{40} (a_{12,j} X_{2,t-j})^2.$$

(8)

滞后 j 年项在总的滞后项中所占的比例如下:

$$\text{countPart}(j) = \frac{\sum_{t=10}^{40} (a_{12,j} X_{2,t-j})^2}{\sum_{j=1}^9 \sum_{t=10}^{40} (a_{12,j} X_{2,t-j})^2}.$$

(9)

通过式(9) 的计算得到,在阶数为 9 时,每 1 项所占的比例分别为:2%,1%, < 1%,7%,22%,7%, < 1%,31%,29%,其中第 8 项所占的比例最大,可见在广州市吸烟率常年来没有显著变化的情况下,广州市的肺癌的死亡率与气溶胶光学消光系数也就是灰霾有 8 年的因果滞后关系.

2 结束语

由于无法对其它省或城市获取相应数据,本文只对广州市的灰霾与肺癌死亡率的情况进行了分析. 新型因果关系法的运用证实了在广州市吸烟率多年来没有显著变化的情况下,肺癌的死亡率与灰霾存在 8 年的因果滞后关系. 该结论与文献[5]得出的结果是一致的. 中国经济的高速发展伴随着自然环境的严重破坏,人们的健康正在遭受着来自恶化的大自然的威胁. 本文的结论无疑告诫人们治理日趋恶化的灰霾环境的必要性和紧迫性.

参考文献

[1] GRANGER C W J. Investigating causal relations by econometric models and cross-spectral methods[J]. *Econometrica*, 1969, 37(37):424-38.

[2] HU S Q, DAI G G, WORRELL G A, et al. Causality analysis of neural connectivity: critical examination of existing methods and advances of new methods[J]. *Neural Networks, IEEE Transactions on*, 2011, 22(6):829-844.

[3] HU S Q, WANG H, ZHANG J H, et al. Comparison Analysis: Granger Causality and New Causality and Their Applications to Motor Imagery[J]. *Neural Networks & Learning Systems, IEEE Transactions on*, 2015, 27(7): 1429-1444.

[4] HU S Q, JIA X X, ZHANG J H, et al. Shortcomings/Limitations of Blockwise Granger Causality and Advances of Blockwise New Causality[J]. *Neural Networks & Learning Systems, IEEE Transactions on*, 2015:1-14.

[5] TIE X X, WU D, BRASSEUR G. Lung cancer mortality and exposure to atmospheric aerosol particles in Guangzhou, China[J]. *Atmospheric Environment*, 2009, 43(14):2375-2377.

[6] ZEIN H, TRAN L H, AZMY A, et al. How to Extract Data from Graphs using Plot Digitizer or Getdata Graph Digitizer[J]. *Technical Report*, 2015(6):1-13.

[7] AKAIKE H. A new look at the statistical model identification[J]. *Automatic Control, IEEE Transactions on*, 1974, 19(6):716-723.

[8] 白雪梅,赵松山. 协整及误差修正模型[J]. *数量经济技术经济研究*,1998(8):39-42.

Causality Analysis from Grey Haze to Lung Cancer in Guangzhou

LIU Xiaoxue, HU Sanqing

(School of Computer, Hangzhou Dianzi University, Hangzhou Zhejiang 310018, China)

Abstract: This paper studies causal relationship from grey haze to lung cancer and further identifies the lag relationship between two variables in Guangzhou city. Firstly, establishing a linear regression model of grey haze and lung cancer mortality; secondly calculating Granger causality value and new causality value between two variables to confirm whether there is causality between them. Thirdly, taking significance test for the two causality values to further verify the causality values are of significance. Finally, calculating different lag part values and comparing the proportion of different lag part takes to confirm the lag relationship between lung cancer mortality and grey haze. With the constructing of linear regression model and the application of causality analysis methods, we can draw a conclusion that grey haze is a causal cause of the rising trend of lung cancer mortality and on average lung cancer mortality and aerosol particles has 8 years lag relationship.

Key words: new causality; Granger causality; lung cancer; grey haze; air pollution