Correlation between Cancer Incidence and Regional Environmental Characteristics Based on Statistical Analysis

Jianqin Sun[a], Yuhan Li[b]

School of Statistics, Shanxi University of Finance and Economics, Taiyuan, Shanxi, China

[a]907399246 @qq.com, [b]807746663@qq.com

Keywords: Environment and Cancer Incidence; Cluster Analysis; Odds Ratio; Interpolation and Fitting; The Fuzzy Set.

Abstract: In recent years, the emergence of "cancer villages" has attracted wide attention from scholars. Then, is there a definite correlation between cancer incidence and regional distribution? Based on the data in the 2017 annual report of tumor registration, this paper made an in-depth study of the influencing factors of cancer incidence through clustering analysis, and found that the municipal cancer incidence areas could be divided into four categories, and proved that cancer incidence and mortality were correlated with the regions. In addition, this paper further was needing to examine the relationship between various regional factors and cancer incidence. In this paper, OR indicated that water quality factors (ammonia nitrogen emission) were significantly corelated with cancer incidence. Through multiple regression analysis and interpolation fitting, the paper concluded that the regional emission of sulfur dioxide, nitrogen oxide emissions, industrial smoke (dust) emissions and the incidence of cancer between the weak correlation.

1. Introduction

Depending on the cancer atlas jointly released by international research institutes, according to the current trend, the number of cancer cases worldwide will increase by 60% by 2040, Lianhe Zaobao of Singapore reported. Among them, smoking, infection, overweight, drinking, workplace carcinogens and outdoor pollution were the principal factors that affected the increase in cancer cases. So are cancer cases related to the quality of the environment in different areas? In this paper, cluster analysis, multiple regression analysis and OR values were used for further study.

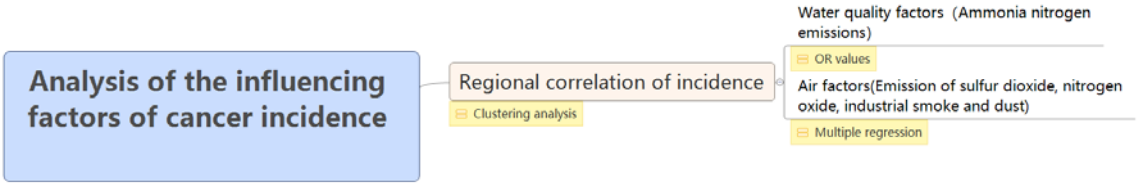

Figure 1. Analysis of the full text.

2. Analysis of regional influencing factors for cancer incidence

From the 2017 tumor registration annual report [1], this paper summarized the number of cancer cases in each registered area (above the urban area) in 2014, and obtained the population of the registered area in 2014 from the 2015 urban statistical yearbook [2], and the incidence formula of the registered area can be obtained as follows:

$$m_{ij} = \frac{n_{m_{ij}}}{N_i}, \quad i = 1, 2, 3\cdots\cdots \tag{1}$$

Preliminary analysis of the data demonstrates that the incidence of oral and pharyngeal diseases varies from region to region. In Genghe city, Inner Mongolia, the incidence of oral and pharyngeal

  DOI: 10.25236/isbsbh.2020.027

diseases is 0.59%, while that of Xilinhot city is as high as 0.34%. The incidence of other parts can be also taken into account in the data and also exists certain relevance between regions.

Table 1. Description of symbols in the model.

Symbol	Symbol Description
m_{ij}	The incidence of the disease in part j of region i
$n_{m_{ij}}$	Number of cases in region i and region j
N_i	Population of area i
C_i	City i
d_{ij}	The distance between two variables
S_j	Variance of j tumor incidence

From the chart, it can be seen that the environment affect the incidence and mortality of tumors to a certain extent, and some scholars have studied the impact of the environment on the incidence of tumors from the perspective of epidemiology [3]. In order to further explore, this paper adopts k-means clustering analysis method to further divide regions.

By using SPSS software, clustering results are obtained as follows:

Category I: Beijing, Tianjin, Shenyang, Shanghai, Hangzhou, Wuhan, Guangzhou.

Category ii: Shijiazhuang, Xinji, Qianan, Qinhuangdao, Wu'an, Baoding, Anguo, Cangzhou, Yangquan, Chifeng, Arun Banner, Genhe, Xilinhot, Zhuanghe, Benxi, Dandong, Donggang, Yingkou, Fuxin, Dehui, Tonghua, Meihekou, Dunhua, Yanji, Shangzhi, Mudanjiang and so on.

Category 3: Dalian, Anshan, Jilin, Wuxi, Changzhou, Suzhou, Nantong, Hefei, Jinan, Qingdao, Yantai, Zhaoyuan.

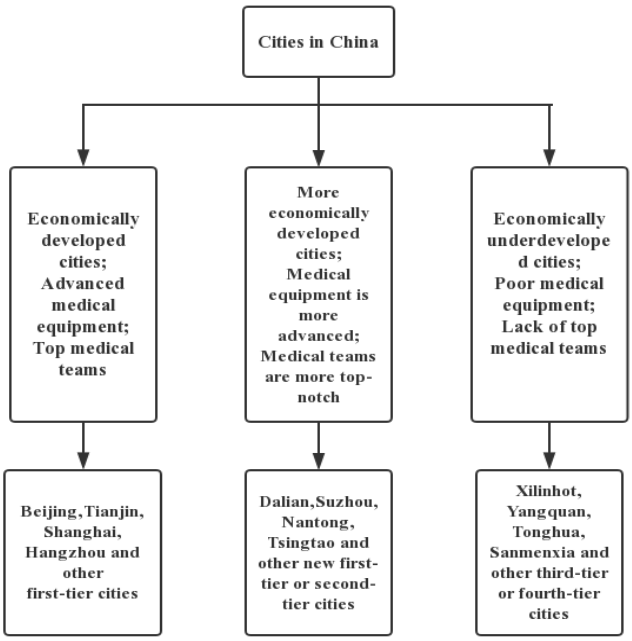

Figure 2. Incidence and Regional Clustering.

The research results in this paper are consistent with pertinent literature [4]. In the report will release by the national cancer center in 2018, the national cancer statistics showed that the incidence of malignant tumors in China ranged from high to low in the east, middle and west. The mortality rate

was highest in the central region than in the eastern and western regions. The high incidence of gastric cancer is concentrated in northwest and coastal provinces and cities, while there has been a high incidence of liver cancer in the southeast coastal region and northeast area, Jilin. Besides, the incidence of esophageal cancer in Hebei and other central plains is also high [5].

3. The relationship between cancer incidence and environmental factors

3.1 The relationship between cancer incidence and water quality factors

According to the cluster analysis, the incidence and mortality of tumors are related to different regions, but the medical conditions and environment of different regions are quite distinct. Here, environmental indicators: ammonia nitrogen emissions, sulfur dioxide emissions, nitrogen oxide emissions, industrial smoke (powder) dust emissions [6] were chosen to study the impact of environment on tumor incidence. The data sources are the urban statistical yearbook 2015, 2015 national statistical yearbook [7]. Due to partial differences in ecological indicators recorded in diverse tumor registration areas, some data are missing. In view of this phenomenon, this paper divides different environmental indicators into water quality factors and is factors and explains their impact on cancer incidence

Firstly, the ecological index of ammonia nitrogen (water quality factor) was tested. This paper sets the collected data and establishes a fuzzy set. Set theory domain U = {Beijing (14000) Tianjin(19164), Shijiazhuang(6425), Shenyang(13027), Jilin(6338.41), Tonghua(3161.7), Mudanjiang(4893.3), Shanghai(41236), Wuxi(3697), Xuzhou(13382), Hangzhou (9058), Hefei (6655), Jinan (5255), Wuhan (13093), Guangzhou (18939), Xining (4050)} (unit: tons).

The membership function $A(x)$ of a fuzzy set "high ammonia nitrogen emission" (A) on U can be defined as:

$$A(x) = \frac{x - 3161.7}{41236 - 3161.7} \tag{2}$$

After establishing fuzzy sets, the median of $A(x)$ is 0.09175, to avoid excessive exposure group (ammonia nitrogen emissions) and lead to higher incidence of exposure group is affected by the dimension, this paper defined at the time $A(x) > 0.1$, the ammonia nitrogen emissions of the high value of ammonia nitrogen emissions areas.

Since this paper is a case-control study in a retrospective study, and the study objects are the "base group" with disease and the "control group" without disease, the OR is more accurate. Although OR has the risk of overstating RR, it is usually caused by a high incidence (>15%), while the cancer incidence in this study was far less than 15%. Therefore, the use of OR can more accurately account for this relationship between ammonia nitrogen emission and cancer incidence.

Table 2. Population data of the exposed group (high ammonia nitrogen emission) and non-exposed group (low ammonia nitrogen emission)

	Disease (people)	No disease (people)	Total number (person)
Exposed group	147113	61234887	61382000
Non-exposed group	42545	22722655	22765200

The calculation formula of OR is as follows:

$$odds1 = \frac{p_1}{1 - p_1} \quad odds0 = \frac{p_0}{1 - p_0} \tag{3}$$

```
. . csi 147133 61234887 42545 22722655, or

                |  Exposed   Unexposed  |      Total
----------------+-----------------------+-----------
          Cases |   147133    6.12e+07  |   6.14e+07
       Noncases |    42545    2.27e+07  |   2.28e+07
----------------+-----------------------+-----------
          Total |   189678    8.40e+07  |   8.41e+07
                |                       |
           Risk |  .7756988    .7293554  |   .7294599
                |                       |
                |  Point estimate       |  [95% Conf. Interval]
----------------+-----------------------+-----------
Risk difference |    .0463434           |   .0444638    .048223
     Risk ratio |     1.06354           |   1.060966   1.066121
Attr. frac. ex. |    .0597441           |   .0574626     .06202
Attr. frac. pop |    .0001432           |
     Odds ratio |     1.283281          |   1.269497   1.297215  (Cornfield)
----------------+-----------------------+-----------

                                chi2(1) =  2059.59  Pr>chi2 = 0.0000
```

Figure 3. Calculation table of OR

The calculation formula of OR confidence interval is as follows:

$$se\left(\ln(OR)\right)=\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d} \tag{4}$$

$$CI\left(\ln(OR)\right)_{(1-\alpha)}:\ln(OR)\pm u_{\alpha}se\left(\ln(OR)\right)\quad CI(OR):\exp\left[\ln(OR)\pm u_{a}se\left(\ln(OR)\right)\right] \tag{5}$$

The confidence interval between cancer incidence and ammonia nitrogen emission OR obtained from figure 2 is [1.269, 1.297].

Table 3. Strength relationship of OR.

OR		Association Strength
0.9-1.0	1.0-1.1	/
0.7-0.8	1.2-1.4	Weak (the former is negatively correlated and the latter is positively correlated)
0.4-0.6	1.5-2.9	Medium (ibid.)
0.1-0.3	3.0-9.0	Strong (ibid.)
<0.1	10 or more	Very strong (ibid.)

According to figure 2, OR is 1.283281. According to table 3, there is a positive influence relationship between ammonia nitrogen emission and cancer incidence, that is, ammonia nitrogen emission is a risk factor, but its impact degree is relatively weak.

3.2 Relationships between cancer incidence and the air factors

Next, we will analyze the impact of the air factor. By 2017 tumor registration report, this paper get the registration area of the oral cavity and throat cancer, nasopharyngeal carcinoma and other 26 kinds of the onset of number, but the 26 kinds of cancer by air environment factors have different effect, so before the formal regression, this article will every cancer with air factor indexes of multiple regression analysis, get the corresponding R^2, the meaning of the variables and the model is as follows:

$$y_i=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon \tag{6}$$

y_i is the number of patients with type i tumor; x_1 is the nitrogen oxide content in the air (unit: ton); x_2 is sulfur dioxide content in the air (unit: ton); x_3 is industrial smoke emission (unit: ton).

In this paper, it concludes that when R^2 >0.25. This kind of cancer was significantly affected by air factors in the environment. The types of R^2 being shown in table 4.

Table 4. Major cancer sites

Sites	R^2
The liver	0.2562
Cervical cancer	0.2536
Throat	0.2930
Trachea, Bronchi, Lungs	0.3406
Other thorax, Trachea	0.3455
Bone	0.3911

As there are only 16 regions where all three indicators exist, the sample number is too low, which may lead to multicollinearity and heteroscedasticity problems caused by the model. Therefore, in order to supplement an air factor index in the environment of some regions, the data were fitted in this paper, and the fitting formula was obtained as follows:

$$x_1 = -2\times10^{-6} + 2.522x_3 + 3795.1, R^2 = 0.4084 \quad (7)$$

$$x_2 = 686.41x^{0.4404}, R^2 = 0.4422 \quad (8)$$

After fitting values of nitrogen oxide and sulfur dioxide were obtained, the sample data were further expanded and multiple regression analysis was performed for the six tumors. Due to nitrogen oxides of sulfur dioxide and fitting value is calculated according to the industrial dust data, in order to avoid multicollinearity problem, this article will industrial soot removed from the original model, the variable regression equation is obtained through the test of significance, it is concluded that nitrogen oxides, sulfur dioxide, and industrial dust air factors affect the conclusion of cancer incidence, and nitrogen oxides, sulfur dioxide and industrial soot emissions, the greater the number of disease. Surprisingly, however, the regression results showed that the influence of environmental air factors on cancer incidence was not as significant, with almost all the impact coefficients below 0.01, which is not the way most people believe.

Through the relevant data, this paper found that air factors associated with cancer, but not the main factor that cause cancer, have relevant literature by comparing the foreign environmental quality good areas of cancer incidence, points out that the main factors that cause cancer for age, smoking, obesity, eating a healthy diet, living habits and genetic and so on. According to the specific data collected in this paper, the cancer incidence rate in Beijing is 2.488%, the average annual concentration of sulfur dioxide is 22 $\mu g / m^3$, the average annual concentration of insoluble particulate matter is 116 $\mu g / m^3$, and the average annual concentration of nitrogen dioxide is 57 $\mu g / m^3$. In Tonghua city, Jilin province, the incidence rate is as high as 6.9923%. The average annual concentration of nitrogen oxide in the air is 36 $\mu g / m^3$. The average annual concentration of sulfur dioxide is 39 $\mu g / m^3$, and the average annual concentration of pm10 is 88 $\mu g / m^3$. In comparison, the air pollution in Beijing is more serious, but the incidence of cancer is far lower than that in Tonghua city of Jilin province, which shows that air quality factors do not possess a significant influence on the incidence of cancer.

4. Conclusions and Suggestions

4.1 Conclusions

Through cluster analysis, this paper concludes that cancer incidence and mortality are correlated with regions, which can be divided into four categories. Thus suggesting that different environmental factors in different regions are the key factors leading to cancer incidence and mortality. On this basis, the fuzzy set was established and the OR was calculated, indicating that water quality factors could also affect the incidence of cancer in a positive way. By expanding the sample data interpolation and fitting method, by using multiple regression model and the concentration of sulfur dioxide and

nitrogen oxides in air industrial soot content on cancer incidence are same, but the effect was not significant, result may be due to the influence of environmental factors on cancer incidence is subtle, lagged effect, so to study the effect of current is not obvious. There are many other factors that affect cancer incidence and mortality, such as:

(1) There are significant differences in medical level between different regions;

(2) Residents in different regions have different degrees of understanding of cancer;

(3) The living habits and eating habits of residents in different areas are different;

(4) The genes of residents in different regions are different;

(5) Residents have different smoking levels [8].

4.2 Suggestions

4.2.1 Regional perspective

(1) Encourage the establishment of organizations to publicize the knowledge of cancer prevention and cancer prevention, popularize the knowledge of cancer;

(2) Strengthen medical, culture and life exchanges between different regions;

(3) Complete the medical construction in various regions, improve the therapeutic level in various regions, and narrow the existing large medical gap;

(4) Encourage local residents to be involved in supervision and report the found illegal pollutant discharge.

4.2.2 Pollutants in the air and water resources

(1) Develop and use clean energy with low sulfur content and find alternatives to coal, so as to reduce the use of coal.

(2) Strict supervision and control to reduce the emission of pollutants into the atmosphere of enterprises and factories.

(3) Encourage afforestation and increase the green area of the region.

(4) Strict supervision and control to reduce the sewage discharge of enterprises and factories.

(5) Check the content of pollutants in local drinking water irregularly.

(6) Residents should not drink water that may have been infected, but should drink boiled water.

References

[1] Zhang, Jiarui, Zeng, Weihua, Shi, Han. Regional environmental efficiency in China: Analysis based on a regional slack-based measure with environmental undesirable outputs[J]. Ecological Indicators, 71:218-228.

[2] Z. Zhao, J. Wang and Y. Liu, "User Electricity Behavior Analysis Based on K-Means Plus Clustering Algorithm," 2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC), Dalian, China, 2017, pp. 484-487. doi: 10.1109/ICCTEC.2017.00111

[3] Carol L. Williams, Matt Liebman, Jode W. Edwards, et al. Patterns of Regional Yield Stability in Association with Regional Environmental Characteristics[J]. Crop Science, 2008, 48(4):1545-1559.

[4] SONG Yonggang, YU Caifen, ZHANG Yufeng, et al. Geochemical Characteristics of Trace Metals in Sediments of Liaodong Bay Based on Multivariate Statistical Analysis[J]. Research of Environmental Sciences, 2016.

[5] Titov A, Gordov E, Okladnikov I. Information-computational system for storage, search and analytical processing of environmental datasets based on the Semantic Web technologies[J]. 2009, 26(112):4607-4614.

[6] Wang, Xinyi, Ji, Hongying, Wang, Qi, et al. Divisions based on groundwater chemical characteristics and discrimination of water inrush sources in the Pingdingshan coalfield[J]. Environmental Earth Sciences, 75(10):872.