Geographical distribution of liver cancer incidence based on big data analysis

Yuchi Tian¹, Qian Huang², Xiaogang Jiang², Guojie Chen², Chongwei Zheng² and Xiaodong Yang¹,³,⁴,*

¹ Academy of Engineering and Technology, Fudan University, Shanghai, 200433, China
² Dalian Naval Academy, Dalian 116018, China
³ Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China
⁴ Jihua Laboratory, Foshan 528000, China

*Corresponding author. Email: xiaodong.yang@sibet.ac.cn

Abstract. The liver cancer has become the sixth most common cancer and the second leading cause of cancer deaths worldwide. The incidence of liver cancer has characteristics of the geographical distribution, which vary from region to region. Based on the data from Cancer Incidence in Five Continents, this study analyzed the spatial distribution of liver cancer globally and explored the effect of meteorological factors of surface temperature and wind field on incidence of liver cancer, to find the geographical distribution characteristics of liver cancer and the impact of geographical environmental factors on the incidence of liver cancer, which will be of great significance to the epidemiological study of liver cancer and the prevention of liver cancer. The results show that the humid and hot oceanic climate may be conducive to the development of liver cancer. The environmental factors play an important role in the occurrence of liver cancer.

Keywords. Liver cancer; geographical distribution; oceanic parameters

1. Introduction
Cancer is the leading cause of death around the world and is a major public health problem, the global cancer deaths per year are projected to more than 11 million in 2030 [1], and liver cancer is one of the most common cancer [2, 3, 4], which is reported as the sixth most common cancer and the second leading cause of cancer deaths worldwide [5]. Liver cancer is mainly divided into two types: hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC). The HCC comprise about 85% of cases. There are various studies about the main risk factors for HCC, such as hepatitis B virus (HBV) or hepatitis C virus (HCV), aflatoxin-contaminated foodstuffs, heavy alcohol intake, obesity, smoking, and type 2 diabetes [6], but few studies on liver cancer impact factors from the perspective of geographical characteristics.

The main risk factors for HCC are many, and the factor of liver cancer vary from region to region, so the incidence of liver cancer must has some characteristics of the geographical distribution. According to the Global cancer statistics (GLOBOCAN) estimates, about 0.84 million new liver cancer
cases and 0.78 million deaths occurred annually in the world, while the most of cases happened in Eastern and South-Eastern Asia [7]. It shows the incidence of liver cancer has obvious characteristics of the geographical distribution. This paper analyses the relationship between geographic position and the incidence of liver cancer, and explores the influence of geographical environment on the prevalence of liver cancer. We aim to provide a scientific basis for the epidemiological research and prevention of liver cancer.

2. Methods
In this study, the method of regression analysis is utilized to study the liner relationship between the incidence of liver cancer and the geographical factor (latitude and longitude), and to apply the trend-surface analysis to show the geographical distribution characteristics of liver cancer incidence, with the data of geographical location information and liver cancer incidences.

2.1. Data
In order to analyses the relationship between geographic position and the incidence of liver cancer, we collect the age-standardized incidence rates (ASRs) of liver cancer from the Cancer Incidence in Five Continents (CI5) [8], which is a long collaboration between the International Agency for Research on Cancer and the International Association of Cancer Registries, to provide a unique source of cancer incidence data from high-quality population-based cancer registries (regional and national) around the world. The ASRs of liver cancer data used for this study is from 378 registries covering in 68 countries for the registration period 2003-2007 in Volume X of CI5, and it includes ASRs of male and female respectively. The latitude and longitude information about these registries are obtained from Google Earth. The Figure 1 and Figure 2 below shows the distribution of liver cancer incidence rates for male and female respectively at all registries worldwide respectively.

![Figure 1. The distribution of liver cancer incidence rates for male at all registries.](image-url)
The Sea surface temperature (SST) data collect from Met Office Hadley Centre, which is a UK's foremost climate change research centre. The spatial resolution of SST data is $1^\circ \times 1^\circ$ and the spatial range is $89.5^\circ S \sim 89.5^\circ N$, $0.5^\circ E \sim 359.5^\circ E$.

The Sea surface wind field data comes from Physical Oceanography Distributed Active Archive Center (PO.DAAC). The spatial resolution of the data is $0.25^\circ \times 0.25^\circ$, the temporal resolution is 6h, the spatial range is $78.375^\circ S \sim 78.375^\circ N$, $179.875^\circ W \sim 179.875^\circ E$, and the time range is from July 1987 to the present [9]. The data can be downloaded from http://rda.ucar.edu/datasets/ds744.4/data/.

2.2. Correlation analysis
The Pearson correlation coefficient is used to measure the linear association between liver cancer incidence and latitude or longitude, the formula is as follows

$$r = \frac{\Sigma_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma_{i=1}^{n}(X_i - \bar{X})^2 \Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Where $r$ is the Pearson sample correlation coefficient between $X$ (latitude) and $Y$ (longitude); $n$ is the number of observations $X_i$ is the value of $X$ (for $i$th observation); $Y_i$ is the value of $Y$ (for $i$th observation).

The value of the correlation coefficient varies between +1 and -1.

The P-value is calculated to determine the representativeness of the sample correlation coefficient to the overall degree of correlation, and we here set the significance level $\alpha$ as 0.01.

2.3. Trend surface analysis
Trend-surface analysis (TSA) is a form of polynomial regression used in geology, ecology and geography to model the overall distribution of properties throughout space. It can fit the liver cancer global trend of distribution as a surface, which is expressed by polynomial function found through the method of least squares.

The F-value and R-squared are used to test the significance and goodness of fit for the surface. R-squared is equal to the sum of squares due to regression divided by the total sum of squares. The F-value is calculated using the formula as follow

Figure 2. The distribution of liver cancer incidence rates for female at all registries.
\[ F = \left( \frac{SSR/(p-1)}{SSE/(n-p)} \right), \]

where SSR is the sum of squares for regression, SSE is the sum of squares for error; n is the number of observations, p is the number of regression parameters.

Setting the significance level \( \alpha \) as 0.01. If the F-value is greater than the critical value of the F-distribution \( F_\alpha \), this would mean that the distribution pattern of properties can be summarised with this surface function.

3. Results

3.1. The distribution of liver cancer registries
Based on the geographic coordinates of the liver cancer registries, we made a spatial distribution map of these registries, and also showed the level of incidence of each registry, which was implemented by Python. It can be found in Figure 1, 2 that in North America and Europe, registries are very densely established, but many other regions around the world do have fewer establishments, especially in Africa and South America.

3.2. Correlation analysis
Using IBM SPSS statistical software to perform linear regression analysis on the incidence of liver cancer and latitude and longitude, the correlation analysis results are shown in Table 1. For latitude, the Pearson correlation coefficients of male and female are 0.11 and -0.68, respectively, and their P-values are 0.825 and 0.191, which are both greater than 0.01. This demonstrates that the linear correlation between incidence and latitude is weak to both male and female. By contrast, the linear correlation between incidence and longitude is extremely strong, which their correlation coefficients and P-values of male and female are 0.436, 0.434, 0.000, 0.000 respectively.

| Table 1. The analysis of the correlation between the incidence of liver cancer and latitude and longitude. |
|---------------------------------------------------------------|
|                  | Latitude | Longitude |       |       |
|                  | Male     | Female    | Male  | Female|
| Pearson correlation coefficient | 0.11     | -0.68     | 0.436 | 0.434 |
| P-value (2-tailed)       | 0.825    | 0.191     | 0.000 | 0.000 |

3.3. Trend surface analysis
According to the longitude, latitude and incidence of liver cancer each registry, a polynomial equation of degree five in binary is established for trend surface analysis. The polynomial equation regression coefficient value of trend surface for male and female, as listed in Table 2. For the test of fit goodness of trend surface, the \( R^2 \) value of male and female are 0.569 and 0.541 respectively. Their fit goodness are both acceptable. For the test of significance, F-values of male (23.439) and female (20.89) are greater than 4.662 of \( F_{0.01} \), indicating that their trend surface equations are both significant. These two polynomials passed the test, so the trend surfaces can be used to analyze the spatial distribution and change trend of liver cancer incidence. Figure 3 and Figure 4 show the 3D Trend surface of male and female liver cancer incidence respectively.

| Table 2. Regression coefficient value of trend surface equation. |
|---------------------------------------------------------------|
|                  | Male       | Female     |
| Variable          | Coefficient| Variable   | Coefficient |
| \( X \)           | -1.34644506e-01 | \( X \)    | -1.97391806e-02 |
| $Y$         | -1.99046793e-01 | $Y$         | -1.47524507e-01 |
|------------|-----------------|------------|----------------|
| $X^2$      | -9.24528107e-04 | $X^2$      | -5.51651294e-04 |
| $XY$       | 7.19361560e-04  | $XY$       | -9.6130314e-04 |
| $Y^2$      | 6.82156210e-03  | $Y^2$      | 2.66417885e-04 |
| $X^3$      | 2.83422682e-05  | $X^3$      | 5.43880798e-06 |
| $X^2Y$     | 1.01417073e-04  | $X^2Y$     | 3.71093201e-05 |
| $XY^2$     | 1.43294658e-04  | $XY^2$     | 7.08071721e-05 |
| $Y^3$      | 1.18302780e-04  | $Y^3$      | 9.9911872e-05  |
| $X^4$      | 3.86915427e-08  | $X^4$      | 2.68824139e-08 |
| $X^3Y$     | -2.01331356e-08 | $X^3Y$     | 1.48014430e-07 |
| $X^4Y$     | 4.52811970e-07  | $X^4Y$     | 5.31701708e-07 |
| $X^4$      | 8.27963613e-07  | $X^4$      | 4.80584900e-07 |
| $Y^4$      | -3.47217751e-06 | $Y^4$      | -1.34624772e-06|
| $X^5$      | -8.06328750e-10 | $X^5$      | -2.18097517e-10|
| $X^4Y$     | -2.64277241e-09 | $X^4Y$     | -1.5530401e-09 |
| $X^5$      | -8.74207469e-09 | $X^5$      | -5.4352135e-09 |
| $X^2Y^2$   | -3.22292357e-08 | $X^2Y^2$   | -1.51375950e-08|
| $XY^3$     | -4.53823118e-08 | $XY^3$     | -1.87387371e-08|
| $Y^5$      | 6.15962911e-09  | $Y^5$      | 3.98632363e-11 |

**Constant term**  
5.876702763216939  

**Constant term**  
4.72317876138257

**Figure 3.** 3D Trend surface of spatial distribution and change trend of male liver cancer incidence.
Discussion

Figure 3 shows that the global geographic distribution of male liver cancer incidence is mainly in North America, Asia, Europe, and there is a trend toward South America and southern Africa. Among them, Asia is a region with a high incidence of liver cancer in men, especially in the eastern and southeastern coastal areas. It can be seen from Figure 4 that the global incidence of liver cancer in women is lower than that in men. Similar to the spatial distribution of male liver cancer, it is also mainly concentrated in the northern hemisphere. The high incidence area is still in Asia. Although the trend surface distribution of male and female diseases shows a high incidence in the northwest region, it is located in the Pacific Ocean and has no human settlements, so it is an outlier and the analysis of this region can be ignored.

Although there is no obvious linear relationship between the latitude and the incidence of liver cancer in the linear relationship analysis, it does show significant regional differences in the trend surface. This may be because that these registries are still insufficient to represent the actual distribution characteristics of liver cancer, while the trend surface predicts the incidence of liver cancer globally, so it can more truly reflect the changing trend of liver cancer incidence. It can be found from Figure 3 and 4 that the areas with high incidence of liver cancer are roughly in the range of 100°E~140°E, 10°S~40°N, which mainly covers the east of Asia and Southeast Asia, but the total of incidence in these areas accounts for almost 80% of the global prevalence of liver cancer incidence. In reaction to this phenomenon, we think whether the prevalence of liver cancer may be related to the characteristics of geographical environment.
Figure 5. Spatial distribution characteristics of sea surface temperature in global waters, February (left), August (right).

As shown in Figure 3 and 4, the liver cancer incidence of male and female both gradually increase from inland to coastal area, and are mainly concentrated in the coastal area of latitudes within 40°N of the northern hemisphere. Considering the environment in this area is greatly affected by the ocean, such as sea surface temperature and wind field, therefore, we plotted sea surface temperature map (Figure 5) and wind field map (Figure 6) in the global waters for analyzing the relationship between liver cancer incidence and meteorological factors in this area. By comparing and analyzing the spatial distribution characteristics of liver cancer cases (Figure 3 and 4) and sea surface temperature (Figure 5), overall, the sea surface temperature to the north of 40°N is significantly lower than that of low and middle latitudes. This means the temperature decreases with increasing latitude, therefore, we initially deduce that the hot and humid oceanic climate is conducive to the occurrence and development of liver cancer. For wind field (Figure 6), there is no obvious relationship between wind field and liver cancer incidence found.
Due to limited data on the incidence of liver cancer, the correlation between the spatial distribution of liver cancer and temperature distribution obtained in this study is still in the conjecture stage. In the future, it is necessary to collect liver cancer cases in various seasons around the world and systematically analyze the relationship between spatial distribution of liver cancer and the spatial distribution of temperature and humidity in each season to provide scientific basis for the prevention and treatment of liver cancer.

5. Conclusions
In this paper, we studied the relationship between geographic position and the incidence of liver cancer, and explored the influence of geographical environment on the prevalence of liver cancer. Our research results show that the humid and hot oceanic climate may be conducive to the development of liver cancer, which may provide a scientific basis for the epidemiological research and prevention of liver cancer. However, the global liver cancer data collected this time is relatively limited. In the future, it is necessary to collect more global liver cancer data in various seasons, and systematically analyze the correlation between the spatial distribution of liver cancer and the spatial distribution of temperature and humidity in each season.

Acknowledgements
The authors acknowledged the funding from Jihua Laboratory under No. X190131TD190, and PLA Dalian Naval Academy for the funding and Academy of Engineering and Technology in Fudan University for lab facilities support.

References
[1] Mathers C D, Loncar D. Projections of global mortality and burden of disease from 2002 to
2030[J]. Plos med, 2006, 3(11): e442.

[2] Ghoncheh M, Mohamm adian M, Mohamm adian-Hafshejani A, Salehiniya H. The incidence and mortality of colorectal cancer and its relationship with the human development index in asia. Ann Glob Health 2016; 82: 726-737.

[3] Mohamm adi M, Naderi M, Ansari Moghaddam A, Mahdavifar N, Mohamm adian M. Investigation of the Relationship between Breastfeeding and Leukemia in Children. Iran J Ped Hematol Oncol 2018; 8: 97-104

[4] Kuang-Rong W, Xia Yu, Rong-Shou Z, Xia-Biao P, Si-Wei Z, Ming-Fang J, Zhi-Heng l, Zhi-Xiong O, Wan-Qing C. Incidence and mortality of liver cancer in China, 2010. Chin J Cancer 2014; 33: 388-394.

[5] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015; 136: E359-E86.

[6] London WT, Petrick JL, McGlynn KA. Liver cancer. In: Thun MJ, Linet MS, Cerhan JR, Haiman CA, Schottenfeld D, eds. Cancer Epidemiology and Prevention. 4th ed. New York: Oxford University Press; 2018:635-660.

[7] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: a cancer journal for clinicians, 2018, 68(6): 394-424.

[8] Curado M P, Edwards B, Shin H R, et al. Cancer incidence in five continents, Volume IX[M]. IARC Press, International Agency for Research on Cancer, 2007.

[9] Atlas R, Hoffman R N, Ardizzone J, et al. A cross-calibrated, multiplatform ocean surface wind velocity product for meteorological and oceanographic applications [J]. American Meteorological Society, 2011, 92: 157-174.