International Health 2021; **13**: 70–79
doi:10.1093/inthealth/ihaa022 Advance Access publication 1 June 2020

ORIGINAL ARTICLE

Spatial distribution of esophageal cancer mortality in China: a machine learning approach

Yilan Liao[a,*], Chunlin Li[a,b], Changfa Xia[c], Rongshou Zheng[c], Bing Xu[a,b], Hongmei Zeng[c], Siwei Zhang[c], Jinfeng Wang[a] and Wanqing Chen[c]

[a]*State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 10010, China;* [b]*College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China;* [c]*National Office for Cancer Prevention and Control, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China*

*Corresponding author: Tel.: +86-010-64889055; E-mail: liaoyl@lreis.ac.cn

Received 8 February 2020; revised 26 March 2020; editorial decision 29 March 2020; accepted 8 April 2020

Background: Esophageal cancer (EC) is one of the most common cancers, causing many people to die every year worldwide. Accurate estimations of the spatial distribution of EC are essential for effective cancer prevention.

Methods: EC mortality surveillance data covering 964 surveyed counties in China in 2014 and three classes of auxiliary data, including physical condition, living habits and living environment data, were collected. Genetic programming (GP), a hierarchical Bayesian model and sandwich estimation were used to estimate the spatial distribution of female EC mortality. Finally, we evaluated the accuracy of the three mapping methods.

Results: The results show that compared with the root square mean error (RMSE) of the hierarchical Bayesian model at 6.546 and the sandwich estimation at 7.611, the RMSE of GP is the lowest at 5.894. According to the distribution estimated by GP, the mortality of female EC was low in some regions of Northeast China, Northwest China and southern China; in some regions downstream of the Yellow River Basin, north of the Yangtze River in the Yangtze River Basin and in Southwest China, the mortality rate was relatively high.

Conclusions: This paper provides an accurate map of female EC mortality in China. A series of targeted preventive measures can be proposed based on the spatial disparities displayed on the map.

Keywords: cancer mapping, esophageal cancer, genetic programming, prevention and control, spatial distribution.

Introduction

Esophageal cancer (EC) is the ninth most common cancer worldwide and the sixth deadliest cancer, causing almost 0.5 million people to die every year globally according to GLOBOCAN statistics.[1] Approximately 49% of all new cases of EC occur in China. Currently, the incidence and mortality of EC is increasing and substantial work is required to understand the causes of this rapid increase in some countries.[2] Accurate geographic distribution of EC incidence and mortality offers an important reference for epidemiological studies of EC.

It should be noted that there are regional differences in the mortality of EC in China because of different living habits, economic conditions and other complex risk factors. Considering the significant spatial heterogeneity of EC,[3] accurate disease mapping is necessary to summarise the incidence and mortality of EC in a country or region, to identify early abnormal outbreaks and spatial variations and to obtain clues regarding the disease aetiology.[4] Additionally, accurate EC mapping can be used as a reference for the allocation of health service resources, especially the quantity, scale, type and location of health services. Previous studies have also noted the important role of cancer maps in the prevention and control of cancer.[5,6] According to the trends, patterns and regional differences associated with the cancer epidemic, we can quickly identify hot spots or outbreaks of EC and correspondingly adjust the quantity, scale and location of health service resource allocation; eventually, we can effectively prevent and treat cancer and reduce the harmful effects of cancer on humans.

However, epidemiological data on EC vary widely in both coverage and quality among countries and regions around the world, ranging from complete coverage by national cancer registries to population-based registries that cover a part of the country, hospital-based registries or areas with no available data.[3] In cases with incomplete data, inference methods must be used based on the available data to provide the best estimate possible. Therefore, it is important to accurately estimate the spatial distribution of cancer through survey data.

There are various ways of estimating the distribution of cancer mortality/incidence based on spatial statistics: (1) geostatistical models[7] combine spatial autocorrelation techniques to smooth the spatial distribution of cancer mortality. However, an important prerequisite is that cancer mortality has a stationary distribution; (2) spatial heterogeneity interpolation models[8–10] are intended for use with spatially stratified heterogeneous data. The sandwich estimation can fully consider the spatial heterogeneity of cancer mortality by partitioning the study area into homogeneous subareas; and (3) the hierarchical Bayesian model[11,12] considers spatial dependence and uses information from neighbouring regions and the entire geographical region to stabilise estimates based on small, local sample sizes within sectors.[13] The basic objective of these methods is to use cancer mortality (or incidence) surveillance data to establish a relationship model and provide predictions at unmonitored points, thus forming a spatially continuous and smooth cancer mortality (or incidence) map. Before cancer mapping, the geostatistical and spatial heterogeneity interpolation methods both require justifications of the spatial characteristics of cancer mortality that will be mapped. The hierarchical Bayesian model introduces many prior distributions and uses only a linear model to express the relationships between risk factors and cancer mortality (or incidence). However, surveillance data in mainland China are often sparsely distributed. Therefore, it is difficult to determine the spatial distribution characteristics of cancer mortality that will be mapped. Moreover, the influence of risk factors on cancer mortality (or incidence) is complicated.

By contrast, genetic programming (GP) can automatically find complex relationships among data without much prior knowledge. This approach attempts to uncover the intrinsic relationships in a dataset by letting the patterns in the data itself reveal the appropriate models, rather than imposing a model structure that is deemed mathematically tractable from a human perspective.[14] Furthermore, GP is especially useful in domains where the exact form of the solution is not known in advance or an approximate solution is acceptable. Therefore, GP can be used as a viable spatial estimation and mapping technique for cancer mortality and occurrence and thus provide a complement to existing spatial estimation and mapping techniques.

Generally, although EC mortality varies greatly by gender, the spatial distribution is highly similar; for example, where female mortality is high, male mortality is also high. Considering the availability of data, this study demonstrates the spatial distribution of female EC mortality in mainland China in 2014 using GP. Experiments show that GP provides a better estimation effect than other classical methods, such as a hierarchical Bayesian model and sandwich estimation. In addition, EC maps clearly show the significant spatial heterogeneity of the EC mortality distribution, which is conducive to the exploration of risk factors for EC. Early screening for EC can occur in certain areas, and medical staff for treating EC and cancer rehabilitation services should be appropriately allocated in areas with high EC mortality.

Materials and methods

Data

Sufficient and high-quality data are essential for estimating the distribution of EC mortality.[10,15–17] Here, EC mortality surveillance data and auxiliary variable data are the primary data used to estimate the cancer mortality distribution.

Cancer mortality data

In this study, the EC mortality surveillance data for 2014 were derived from the cancer incidence and mortality survey conducted by the National Central Cancer Registry of China.[18] In total, 964 surveyed counties covering almost all 31 provinces and municipalities were included in this study, accounting for 288 million people. The spatial distribution of the counties in the Chinese mainland is shown in Figure 1.

Auxiliary variable data

In this study, the auxiliary variable data were taken from previous literature or widely accepted sources. These variables were then filtered to identify useful variables and remove variables that may bias the estimation accuracy. According to previous studies,[3,19–26] we divided the complicated risk factors into three categories: (1) physical conditions, such as obesity, hormone level and genetic background; (2) living habits, such as eating preferences, living environment and frequency of drinking and smoking; and (3) living environment, such as socioeconomic status and geographical environment. The variables or proxy variable data and the corresponding collection dates are listed in Table 1. Female EC mortality in 1975 was taken from *The Atlas of Cancer Mortality in the People's Republic of China*[27]; socioeconomic data such as the per-capita gross domestic product came from the *China Statistical Yearbook*[28]; demographic data, such as the average life of females, urbanisation rate and the average level of education, were extracted from the Sixth National Population Census (2010)[29]; and health data, such as smoking rates, drinking rates, fruit and vegetable intake and the average level of body mass index (BMI), were extracted from the Nutrition and Health Status of the Chinese People.[30] By regional statistics, the spatial resolution of all variables is county level, which is consistent with EC mortality.

Methodology

GP

We employed a GP method to generate potentially useful models and to estimate the distribution of female EC mortality. GP is a search and optimisation technique inspired by biological evolution.[31] According to the principles of 'survival competition' and 'survival of the fittest,' the step-by-step solution approach obtains the optimal solution from the initial solution by means of automated generation, replication, exchange, mutation and other operations. We obtained the relationship between the mortality of EC (y) and the risk factors (x_1, x_2, ... x_n) through the surveillance data of female EC mortality and the risk factor data, $y = f(x_1, x_2, ... x_n)$. GP takes the error between the estimated value and actual

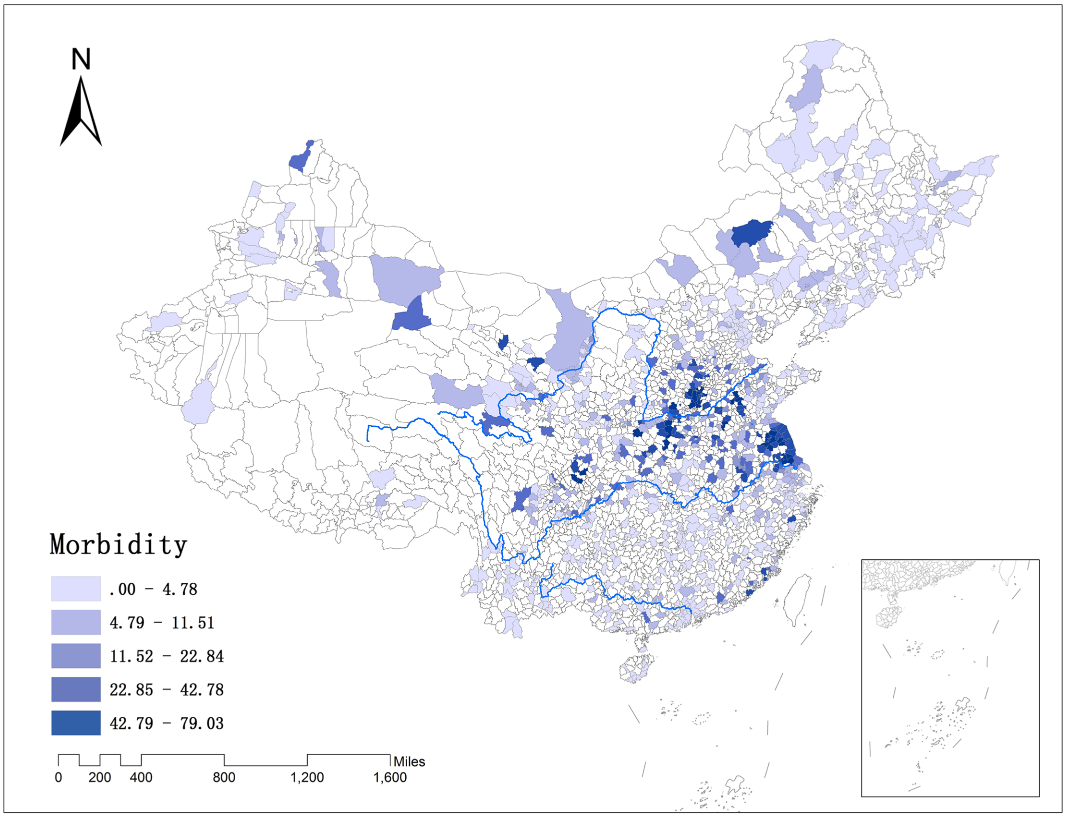

Figure 1. Distribution of female esophageal cancer (EC) surveillance counties in mainland China in 2014 (unit: 1/100 000 females).

value as the driving force of heredity. Then, we used the risk factor data to estimate the mortality of female EC without surveillance data using the most appropriate model obtained. The basic principles and modelling process of GP are described below.

In GP, computer programs are represented in tree structures that are easily evaluated in a recursive manner.[32] Every computer program represents the relationship between the dependent variable (female EC mortality) and the independent variables (the risk factors of female EC). A tree node is a primitive function (+, -, *, /, log, exp, cos, sin, tan....) and a terminal node is a variable (a constant or a risk factor). A GP tree is an individual in a population.

GP first generates a set of naive random individuals to represent the relationship between known risk factors and EC mortality. The initial individuals are a totally random mix of the primitive functions, the risk factors and constants. GP then evaluates the performance of each individual in the population to determine the corresponding fitness level. Fitness is the driving force of natural selection in GP and is basically the same thing as a 'score,' 'error' or 'loss' variable. In this study, fitness is determined by using the root square mean error (RMSE), which is calculated by fitting the surveillance data. Then, individual cases are selected from the population with a probability based on fitness and added to the next generation without making any changes. This study is based on tournaments. In addition, GP involves the selection and recombination of appropriate individual expressions in the population through two major genetic recombination operators: crossover and mutation. In a crossover operation, two individuals are chosen based on fitness and probability. A subtree (or node) is randomly selected for each individual. Two subtrees (or nodes) are exchanged to create two new individuals. In mutation operations, a single individual is chosen randomly as the parent individual. Then, a randomly selected subtree (or node) of the parent individual is replaced by using a new randomly generated subtree (or node) to create a new individual. All the generated new individuals enter the next generation of the population. Crossover and mutation can effectively avoid the population from reaching a global suboptimal solution. Finally, the population is iteratively optimised until the optimal fitness individual is found for a certain generation, which is the optimal solution or approximate optimal solution of the problem. We use this optimal solution as the mapping relationship between independent variables and dependent variables for spatial distribution estimation. The basic flow chart of GP is shown in Figure 2.

Hierarchical Bayesian model

The hierarchical Bayesian model is a popular approach for identifying regions with unusual mortality levels, temporal trends or both.[11] The hierarchical method defines the probability distribution parameters of EC mortality (θ_i) in every region i as a Poisson distribution[33,34] and the Bayesian method is used to estimate the posterior distribution. The model can be divided into three levels as follows.

Table 1. Risk factors and corresponding auxiliary variables

Risk factor	Auxiliary variable	Abbreviation	Year
Physical conditions	EC mortality rate of females	EC 1975	1975
	Average life of females	life female	2014
	Average BMI (females)	BMI	2002
	Rate of overweight females (%)	overweight	2010
	Prevalence of *Helicobacter pylori* infection	HP infection	2010
	Proportion of the minority population (%)	minority	2010
Living habits	Current smoking rate (%) of women (county estimate)	smoke	2002
	Current drinking rate (%) of women (county estimate)	drink	2002
	Rate of excessive red meat intake by women (%)	redmeat	2013
	Rate of insufficient vegetable and fruit intake by women (%)	fruit	2013
Living environment	Urbanisation rate (%)	urban	2014
	Per capita gross national product (trillion dollars)	GDP	2014
	Average level of education for women (year)	edu	2010
	Proportion of the non-agricultural population (%)	nonagri	2010
	Proportion of primary industry (%)	primary industry	2000
	Mean elevation (km)	DEM	2010

Abbreviations: BMI, body mass index; DEM, digital elevation model; EC, esophageal cancer; GDP, gross domestic product; HP, *Helicobacter pylori*.

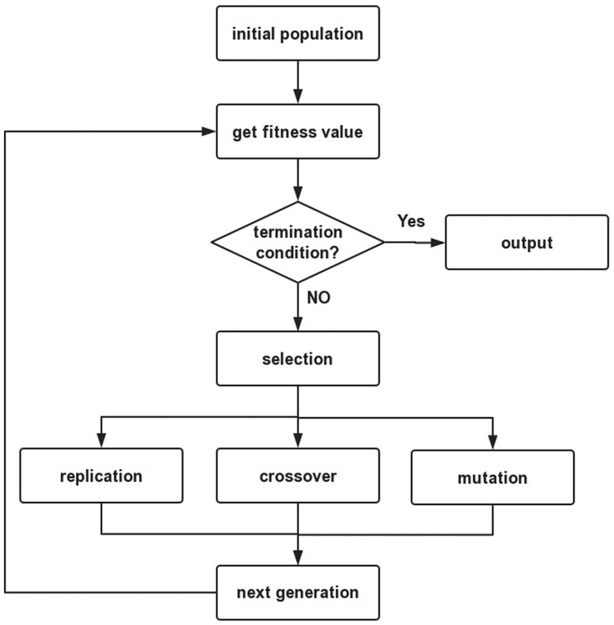

Figure 2. Basic flow chart of genetic programming (GP).

The first level is the Poisson likelihood:

$$Y_i \sim Poisson\ (E_i\theta_i) \tag{1}$$

The model is introduced at the second level:

$$\log(\theta_i) = T + X^T U + \upsilon\ (i) + e\ (i) \tag{2}$$

where θ_i is the relative risk of cancer in region i and E_i is proportional to the total population of region i. The logarithm of the relative risk θ_i is modelled as the mixture of three components: X^T and U denote the auxiliary variables and corresponding coefficient, respectively; $\upsilon(i)$ reflects the spatial structure; and $e(i)$ reflects the heterogeneity of region i.

The third level of the hierarchy consists of the definition of the prior distribution as the indicator of the unknown true status.[33] For example, $e(i)$ is assumed to be a Gaussian distribution, where $e(i) \sim Norm(0,\ \sigma_i^2)$. The variance of each parameter is usually assumed to be subject to gamma (a, b). The method was implemented using R 3.5.0 software and performed using the R package INLA.

Sandwich estimation

Sandwich estimation is a spatial estimation method for multi-unit stratified heterogeneous surfaces.[10] This approach consists of three layers: the sampling layer, the zoning layer and the reporting layer.[9] The sandwich estimation enables accurate multi-unit reporting with few sample points and provides a simple and straightforward way to solve the problem of data transfer between multiple reporting units.[9,10] The sampling layer is a collection of EC mortality surveillance points and the zoning layer divides the study area into multiple areas with the same spatial properties.[9] Zoning layers are risk factors for EC. The reporting layer encompasses the administrative units of counties in this study. Sampling data are passed to the reporting layer through the zoning layer.[10] The data and error streams transfer the estimated mean and sampling error in a layer-by-layer manner. Finally, the results of the estimation of each risk factor are fused based on the variance as a weighted average. Sandwich estimation was performed using R 3.5.0 software and the R package gstat, and the code was written in Sandwich software, Beijing, China.[35]

Table 2. Summary statistics of the female esophageal cancer (EC) mortality surveillance data (n = 964, unit: 1/100 000 females)

Cancer	Mean	Median	Min.	Max.	STD	Moran's I
EC	8.586	4.076	0.000	79.028	148.97	0.691**

$^{**}P < 0.01$.

Evaluating the accuracy of mapping methods

This study compares the effectiveness of GP, the hierarchical Bayesian model[34] and sandwich estimation[9] in estimating the distribution of female EC mortality in mainland China in 2014. The accuracy of the three methods was evaluated using 10-fold cross-validation, which is commonly used for evaluating the accuracy of spatial estimation methods, and the RMSE and coefficient of determination (R^2) were calculated. The female EC mortality point samples were randomly split into 10 subsample sets. Nine subsample observations were used as training models and the remaining subsample was used to test the model. The cross-validation process was then repeated 10 times, with each of the subsamples used exactly once as validation data. To avoid contingency caused by dividing the subset, 10-fold cross-validation was repeated 10 times. A small RMSE reflects a precise spatial estimation method and a large R^2 indicates a precise spatial estimation method.

Results

Description and spatial autocorrelation of female EC mortality data

Table 2 gives a description of the 2014 county-level female EC mortality data. The lowest rate of female EC mortality was 0.000 cases per 100 000 females and the highest rate was 79.028 cases per 100 000 females. The mean and median rates were 8.586 and 4.076 per 100 000 females, respectively (Table 2), which indicated that female EC mortality is characterised by considerable geographical variation. Spatial autocorrelation is commonly used in spatial analysis and refers to features for which the influence associated with an adjacent feature is greater than the distance from that feature. Moran's I is one of the classic statistics used to examine spatial autocorrelation. The data have a Moran's I value of 0.691, which indicates strong spatial autocorrelation.

Pearson correlation coefficient of these auxiliary variables

The Pearson correlation coefficients of these auxiliary variables and female EC mortality were calculated (Table 3). According to the results, the 12 auxiliary variables selected for the calculation were EC 1975, life_female, overweight, HP_infection, minority, smoke, drink, fruit, urban, edu, nonagri and primary industry.

Table 3. Pearson correlation coefficients of auxiliary variables

Auxiliary variable	*r*	Auxiliary variable	*r*
EC 1975	0.822**	redmeat	−0.283**
life_female	−0.052	fruit	−0.170**
BMI	0.072	urban	−0.145**
overweight	0.227**	GDP	−0.046
HP_infection	0.135**	edu	−0.139**
minority	−0.198**	nonagri	−0.186**
smoke	−0.215**	primary industry	0.171**
drink	−0.234**	DEM	−0.124*

Abbreviations: BMI, body mass index; DEM, digital elevation model; EC, esophageal cancer; GDP, gross domestic product; HP, *Helicobacter pylori*.
$^{*}P < 0.05$; $^{**}P < 0.01$.

Table 4. Accuracy comparison of the three methods

Method	RMSE	R^2
GP	5.894	0.767
Hierarchical Bayesian model	6.546	0.712
Sandwich estimation	7.611	0.618

Abbreviations: GP, genetic programming; RMSE, root square mean error; R^2, coefficient of determination.

Comparison of the three spatial estimation methods

The model for estimating the mortality of female EC by GP is shown below.

$$\sqrt{\sqrt{\sqrt{EC\ 1975}+edu+drink+EC\ 1975}/EC\ 1975+overweight}\ /edu+\sqrt{EC\ 1975} \tag{3}$$

Here, the GP selected four risk factors among the 13 risk factors entered: the female EC mortality in 1975, average level of education for women, rate of overweight females and current drinking rate of women. The selected risk factors are the most important risk factors to various extents.

Moreover, we used two other commonly used mapping methods, the hierarchical Bayesian model and sandwich estimation, to estimate the spatial distribution of female EC and compared the accuracy and reliability of the estimation results of the three methods. The GP, hierarchical Bayesian model and sandwich estimation used the same filtered auxiliary variables. The 10-fold cross-validation accuracy comparison for the three methods is shown in Table 4. The RMSE values of the three methods are 5.894 for GP, 6.546 for the hierarchical Bayesian model and 7.611 for sandwich estimation, indicating that GP was the best method, followed by the hierarchical Bayesian model and sandwich estimation. The R^2 values ranked the methods in the same order as the RMSE. In terms of precision, GP is more accurate than the

Table 5. Statistics of the estimations with the three methods (morbidity unit:1/100 000 females)

EC Morbidity	Surveillance data	GP	Hierarchical Bayesian model	Sandwich estimation
<5.5	568 (58.92%)	1630 (57.17%)	1458 (51.14%)	1659 (58.19%)
5.5–12.5	215 (22.30%)	700 (24.55%)	903 (31.67%)	658 (23.08%)
12.5–23.5	89 (9.23%)	342 (12.00%)	324 (11.37%)	426 (14.94%)
23.5–41.5	56 (5.81%)	126 (4.42%)	122 (4.28%)	72 (2.53%)
>41.5	36 (3.73%)	53 (1.86%)	44 (1.54%)	36 (1.26%)
Total	964 (100%)	2851(100%)	2851 (100%)	2851 (100%)

Abbreviations: EC, esophageal cancer; GP, genetic programming.

other two methods in estimating the distribution of female EC mortality in mainland China in 2014.

The distribution of the percentage differences between the estimated mortality of the surveillance data and the three methods is illustrated in Table 5. In the surveillance data, the female EC mortality of most counties (58.92%) was between 0 and 5.5 per 100 000 females. The estimated female EC mortality of most counties (57.17% for GP and 58.19% for sandwich estimation) was between 0 and 5.5 per 100 000 females, but this value was always less than 5 based on the hierarchical Bayesian model (51.14%); this result is smaller than the survey value and those of the other two methods of estimation and there are some negative values that are obviously not reflective of the actual mortality rate. The estimated female EC mortality at the county level (31.67%) varied between 5.5 and 12.5 per 100 000 females for the hierarchical Bayesian model and this range was greater than those of the other two methods and the survey data. In addition, the estimated female EC mortality at the county level (1.86%, 1.54% and 1.26%) was generally more than 41.5 per 100 000 females for the three methods, which was less than the surveyed value (3.73%). This result indicates that the three methods underestimated the value at this scale.

Moreover, we calculate the absolute error (the estimated value minus the monitored value) of the three methods to compare the estimated deviations of the three methods. The three estimated distributions of the absolute error are shown in Figure 3. GP estimates have large deviations downstream of the Yellow River Basin, with both overestimated and underestimated values, but compared with Figure 1 the area downstream of the Yellow River Basin has a high surveyed mortality, so the relative error in this region is small. Additionally, the deviation in the GP estimates is less than 6 (morbidity, 1/100 000) in other regions of the country. The underestimation of the hierarchical Bayesian model results is mainly concentrated downstream of the Yellow River Basin. The overestimated areas are relatively discrete and distributed downstream of the Yellow River Basin and in southern China. The areas with underestimated values based on sandwich estimation are concentrated in the downstream region of the Yellow River Basin, with overestimated areas in northeastern China, southern China and central China. In summary, the results of GP were closest to the actual female EC mortality data.

The three estimated distributions of female EC mortality in the Chinese mainland in 2014 are shown in Figure 4. By a comparison with Figure 1 we can evaluate each method. The results of the three methods basically agree with the distribution trend of the survey data. The mortality of female EC estimated by the three methods is low in northeastern China, northwestern China and southern China and high in the downstream region of the Yellow River Basin. GP and the hierarchical Bayesian model also predict high values in areas north of the Yangtze River in the Yangtze River Basin, Southwest China and the northeastern Xinjiang Province. EC mortality was highest in Hebei Province, Shanxi Province, Henan Province and Jiangsu Province. GP estimated EC mortality is low in Northeast China, Tibet and South China. To sum up, in comparison to the distribution of survey data, the GP, hierarchical Bayesian model and sandwich estimation all provided reliable estimates of the distribution trend.

Discussion

In this study, we used three methods—GP, hierarchical Bayesian model and sandwich estimation—to estimate the spatial distribution of female EC mortality in mainland China in 2014 based on survey data and risk factor data. The hierarchical Bayesian model considers the spatial autocorrelation of adjacent regions and it can manage the uncertainty of sampling data using a framework of probability theory. Sandwich estimation considers the heterogeneity of female EC mortality by zoning a study area into homogeneous subareas. GP can effectively reveal the relationship between female EC mortality and risk factors. For female EC mortality in the Chinese mainland, the spatial distribution of GP estimates is better than those based on the other two methods.

GP has been successfully applied for the spatial estimation of population[36] and climate data.[37,38] This paper verifies that GP has considerable advantages and broad application prospects in estimating the spatial distribution of EC. First, GP can screen for risk factors that do not require prescreening through prior knowledge of cancer pathogenesis. When determining the risk factors for cancer, all the risk factors that may be related to a certain cancer type can be used, without the need to be extremely precise. Although GP can filter variables, to ensure that the variables input by the three methods are consistent and convenient for precision comparison, all three methods use the same filtered variables. Second, the relationship between EC mortality and risk factors is complex. GP is a machine learning approach, which

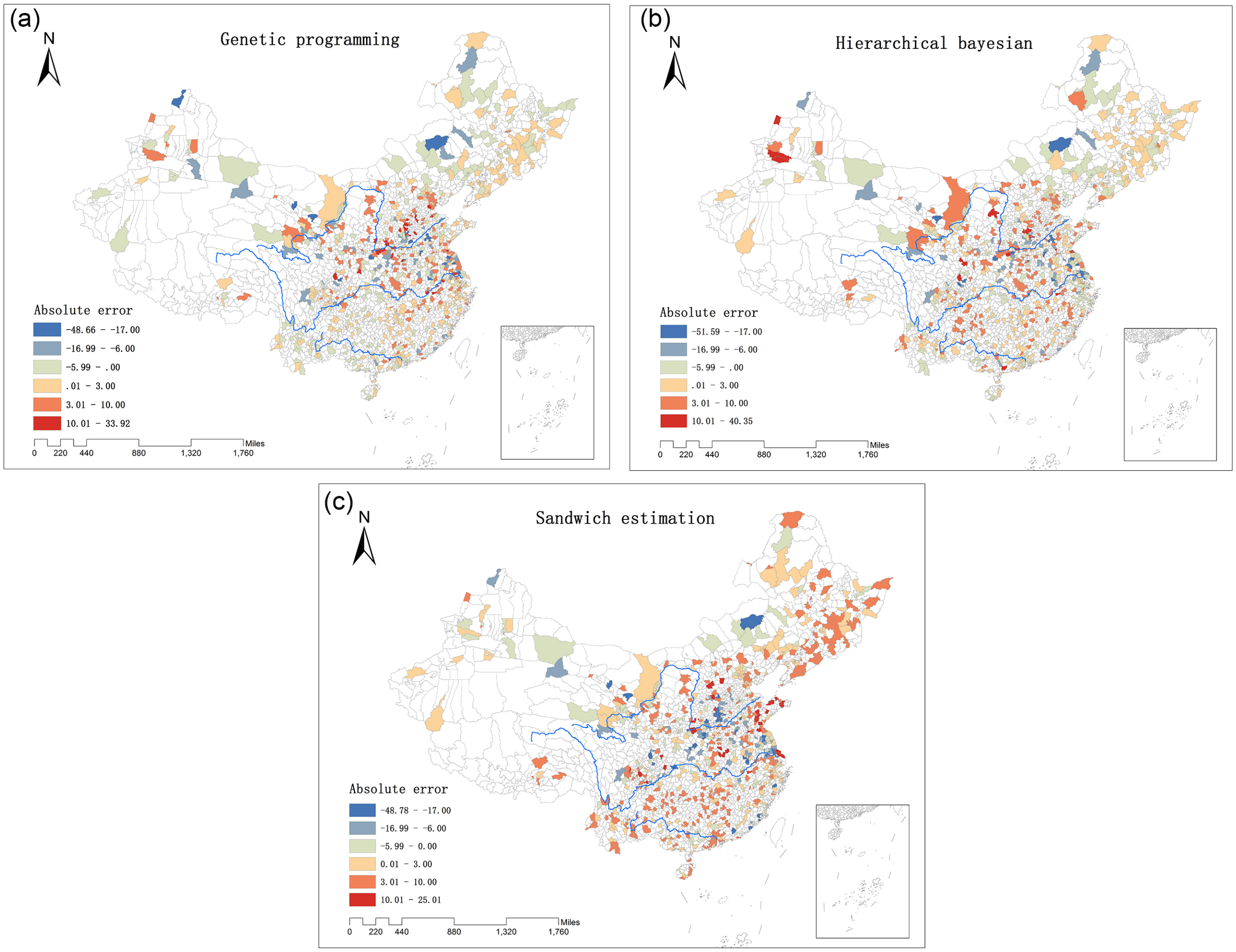

Figure 3. (A) Distribution of the absolute error of female esophageal cancer (EC) mortality estimated by genetic programming (GP) (unit: 1/100 000 females). (B) Distribution of the absolute error of female EC mortality estimated by the hierarchical Bayesian model (unit: 1/100 000 females). (C) Distribution of the absolute error of female EC mortality estimated by sandwich estimation (unit: 1/100 000 females).

can fully explore the relationship between EC mortality data and risk factor data. This approach attempts to reveal the internal relationships in the dataset, rather than imposing a model structure from a human perspective. Most existing methods require some assumptions. These assumptions greatly limit the flexibility of these methods for estimating the spatial distribution of different cancers. If the assumptions are not fully met, the results may be poor. For example, kriging requires a strong spatial autocorrelation among cancer mortality, sandwich estimation requires cancer mortality data with high spatial heterogeneity and the hierarchical Bayesian model requires spatial distribution models that satisfy certain statistical distributions. GP is a data-driven approach that does not require strict assumptions, which makes it less restrictive when estimating the spatial distribution of cancer. Finally, GP estimates the spatial distribution to obtain a relationship between cancer mortality and risk factors. According to the expression of the relationship between the risk factors selected by the model and the mortality of cancer, it is possible to perform prevention and control measures in a targeted manner.

The spatial distribution of female EC in the Chinese mainland was reliably estimated by GP. Based on the map of female EC mortality, some guidelines for the prevention and control measures are proposed. The high-value areas of female EC mortality include the downstream region of the Yellow River Basin, the area north of the Yangtze River in the Yangtze River Basin, Southwest China and the northeastern Xinjiang Province, and these findings are consistent with the results of other studies.[39,40] This approach will enable the government to better target health interventions and efficiently choose areas in which to implement control measures and avoid female EC. Early screening measures for EC, medical staff for treating EC, essential medical services and cancer rehabilitation services should mainly be deployed in areas with high EC mortality. Disease prevention and control organisations should collect accurate hospital-based data from different disciplines and centres in high-mortality areas and investigate the differences among risk factors to strengthen prevention and improve publicity.

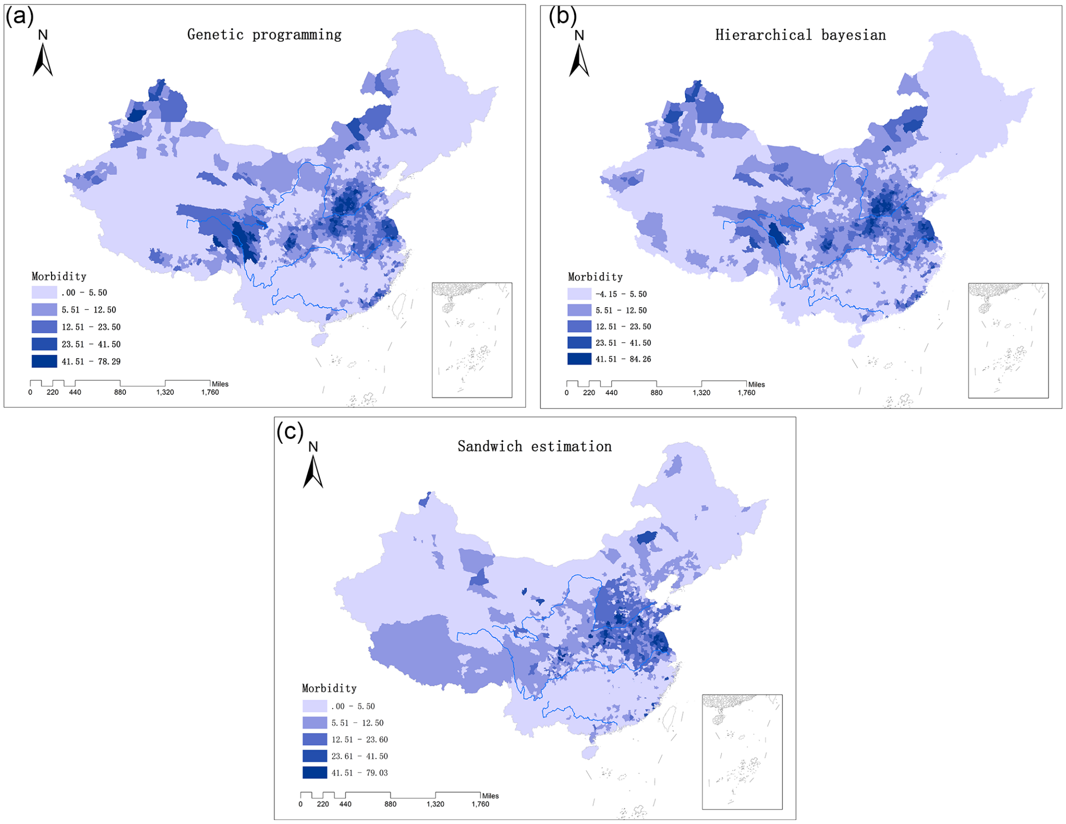

Figure 4. (A) Distribution of the estimated female esophageal cancer (EC) mortality in the Chinese mainland in 2014 by genetic programming (GP) (unit: 1/100 000 females). (B) Distribution of the estimated female EC mortality in the Chinese mainland in 2014 by the hierarchical Bayesian model (unit: 1/100 000 females). (C) Distribution of the estimated female EC mortality in the Chinese mainland in 2014 by sandwich estimation (unit: 1/100 000 females).

Moreover, through the model of risk factors and female EC mortality based on GP, we can propose some recommendations. First, the healthcare system could be improved and the burden of EC screening and treatment could be reduced through financial support. For example, individuals with long-standing and severe gastro-esophageal reflux disease symptoms have a 40-fold increased risk of being diagnosed with EC.[26] The timely and rational allocation of medical and health service resources for some diseases can decrease the risk of EC. Second, the risk factors selected by GP indicate that genetics and the family history of disease are important factors related to the spatial disparities in female EC mortality. Additionally, the level of education is associated with differences in dietary and health ideologies, which leads to different female EC risks. Moreover, women who are overweight and drink excessively are more likely to be diagnosed with female EC. Therefore, prevention and control of female EC should involve increasing public awareness, encouraging physical activity and obesity reduction and promoting reduced drinking.

Nonetheless, there are some limitations to this study. First, the study only focused on female EC because of the lack of data for males. However, past studies have shown that EC mortality was higher in males than in females. We will estimate the spatial distribution of male EC mortality and compare the results in future studies when male EC data can be obtained. Second, although we did our best to find variable data that matched the time of EC mortality, there were still data for some variables which were difficult to obtain. The data of variables were collected in different years and this may have impacted the results. Third, GP is a data-driven method that has certain requirements regarding the amount of data and is prone to overfitting and underfitting problems. In addition, GP has other limitations, such as requiring many hyperparameters that need to be appropriately set and high computational costs. These issues will all be explored in future work.

Conclusions

Accurate cancer maps are important for identifying spatial disparities in cancer and trends in cancers. GP can be used as a viable spatial estimation and mapping technique for cancer mortality and incidence. This approach can complement existing spatial estimation and mapping techniques. This method avoids strict assumptions and provides strong robustness and high precision. In addition, this paper provides an accurate map of female EC mortality in China. According to the spatial disparity displayed on the map, a series of targeted prevention measures are proposed.

Authors' contributions: The authors wish it to be known that, in their opinion, YL and CL should be regarded as joint first authors. YL developed the research concept and revised the manuscript. CL performed experiments with the three methods and wrote the manuscript. YL and CL contributed equally to this work. BX revised the manuscript. CX, RZ, HZ, SZ, JW and WC contributed to the acquisition and interpretation of data. All authors read and approved the final manuscript.

Acknowledgements: Not applicable.

Funding: The authors disclose receipt of the following financial support for the research, authorship and/or publication of this article. This work was supported by the National Key R&D Program of China [grant numbers 2016YFC1302504], the National Natural Science Foundation of China [grant numbers 41471377, 41531179 and 41421001], and a Program Grant in Fundamental Research from the Ministry of Science and Technology [grant number 2014FY121100].

Competing interests: None declared.

Ethical approval: Not required.

References

1 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. CA Cancer J Clin. 2018;68(6):394–424.

2 Pennathur A, Gibson MK, Jobe BA, et al. Oesophageal carcinoma. Lancet. 2013;381(9864):400–12.

3 Stewart B, Wild CP. World cancer report 2014. 2014.

4 Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res. 2005;14(1):35–59.

5 d'Onofrio A, Mazzetta C, Robertson C, et al. Maps and atlases of cancer mortality: a review of a useful tool to trigger new questions. Ecancermedicalscience. 2016;10:670.

6 Ezimand K, Abdolahi Kakroodi A, Javanbakht M. Geographic distribution and incidence of skin cancer using the Geographically Weighted Regression model. J Dermatol Cosmet. 2018;9(1):35–45.

7 Dhaher GM, Lee MH. A comparison between the performance of kriging and cokriging in spatial estimation with application. Matematika. 2013;29:33–41.

8 Hu Y, Bergquist R, Lynn H, et al. Sandwich mapping of schistosomiasis risk in Anhui Province, China. Geospat Health. 2015;10(1):324.

9 Wang J-F, Haining R, Liu T-J, et al. Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface. Environ Plan A. 2013;45(10):2515–34.

10 Liao Y, Li D, Zhang N, et al. Application of sandwich spatial estimation method in cancer mapping: A case study for breast cancer mortality in the Chinese mainland, 2005. Stat Methods Med Res. 2019;28(12):3609–26.

11 Schrödle B, Held, L. Spatio-temporal disease mapping using INLA. Environmetrics. 2011;22(6):725–34.

12 Johnson GD. Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling. Int J Health Geogr. 2004;3(1):29.

13 Achia TN. Spatial modelling and mapping of female genital mutilation in Kenya. BMC Public Health. 2014;14(1):276.

14 Rainville F-MD, Fortin F-A, Gardner M-A. symbolic regression. Available from https://deap.readthedocs.io/en/0.7-0/examples/symbreg.html [accessed July 14, 2019].

15 Li J, Heap AD. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. Ecol Inform. 2011;6(3-4):228–41.

16 Yao X, Fu B, Lü Y, et al. Comparison of four spatial interpolation methods for estimating soil moisture in a complex terrain catchment. PLoS One. 2013;8(1):e54660.

17 Azpurua MA, Ramos KD. A comparison of spatial interpolation methods for estimation of average electromagnetic field magnitude. Prog Electromagn Res. 2010;14:135–45.

18 Chen W, Sun K, Zheng R, et al. Cancer incidence and mortality in China, 2014. Chinese J Cancer Res. 2018;30(1):1.

19 Aghcheli K, Marjani H-A, Nasrollahzadeh D, et al. Prognostic factors for esophageal squamous cell carcinoma—a population-based study in Golestan Province, Iran, a high incidence area. PLoS One. 2011;6(7):e22152.

20 Pandeya N, Olsen CM, Whiteman DC. Sex differences in the proportion of esophageal squamous cell carcinoma cases attributable to tobacco smoking and alcohol consumption. Cancer Epidemiol. 2013;37(5):579–84.

21 Morita M, Kumashiro R, Kubo N, et al. Alcohol drinking, cigarette smoking, and the development of squamous cell carcinoma of the esophagus: epidemiology, clinical findings, and prevention. Int J Clin Oncol. 2010;15(2):126–34.

22 Wu C, Kraft P, Zhai K, et al. Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. Nat Genet. 2012;44(10):1090.

23 Hirota T, Takahashi A, Kubo M, et al. Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. Nat Genet. 2011;43(9):893.

24 Wu M, van't Veer P, Zhang Z-F, et al. A large proportion of esophageal cancer cases and the incidence difference between regions are attributable to lifestyle risk factors in China. Cancer Lett. 2011;308(2):189–96.

25 Wang J-B, Fan J-H, Liang H, et al. Attributable causes of esophageal cancer incidence and mortality in China. PLoS One. 2012;7(8):e42281.

26 Lagergren J, Bergström R, Lindgren A, et al. Symptomatic gastroesophageal reflux as a risk factor for esophageal adenocarcinoma. N Engl J Med. 1999;340(11):825–31.

27 Editorial Committee of the People's Republic of China Malignant Tumor Atlas. Atlas of Cancer Mortality in the People's Republic of China. Beijing: China Map Press; 1979.

28 House CSP. China Statistical Yearbook. Beijing: National Bureau of Statistics of China; 2014.

29 China NBoSo. The Sixth National Population Census. Available from http://www.stats.gov.cn/english/statisticaldata/censusdata/ [accessed July 14, 2019].

30 Ministry of Health of the People's Republic of China. *The Nutrition and Health Status of the Chinese People*. Beijing: People's Medical Publishing House; 2008.

31 Affenzeller M, Wagner S, Winkler S, et al. Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications. Boca Raton, Florida: Chapman and Hall/CRC; 2009.

32 Cramer NL; A representation for the adaptive generation of simple sequential programs. Proceedings of the First International Conference on Genetic Algorithms, 1985, 183–7.

33 Catelan D, Lagazio C, Biggeri A. A hierarchical Bayesian approach to multiple testing in disease mapping. Biometrical J. 2010;52(6):784–97.

34 Haining RP, Haining R. Spatial Data Analysis: Theory and Practice. Cambridge: Cambridge University Press; 2003.

35 Wang J. Sandwich Spatial Estimation. Available from http://www.sssampling.cn/sandwich/ [accessed July 14, 2019].

36 Liao Y, Wang J, Meng B, et al. Integration of GP and GA for mapping population distribution. Int J Geogr Inf Sci. 2010;24(1):47–67.

37 Adhikary SK, Muttil N, Yilmaz AG. Genetic programming-based ordinary kriging for spatial interpolation of rainfall. J Hydrol Eng. 2015;21(2):04015062.

38 Kumar Adhikary S, Muttil N, Gokhan Yilmaz A. Ordinary kriging and genetic programming for spatial estimation of rainfall in the Middle Yarra River catchment, Australia. Hydrol Res. 2016;47(6):1182–97.

39 He Y-T, Hou J, Chen Z-F, et al. Trends in incidence of esophageal and gastric cardia cancer in high-risk areas in China. Eur J Cancer Prev. 2008;17(2):71–6.

40 Hao W. Spatial Distribution of Hepatic Cancer and Esophageal Cancer Mortality in Shandong Province from 2011 to 2013. Shandong: Shan Dong University; 2015.