SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Review Article

Statistical Methods in Medical Research
0(0) 1–18
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280218811344
journals.sagepub.com/home/smm
SAGE

Application of sandwich spatial estimation method in cancer mapping: A case study for breast cancer mortality in the Chinese mainland, 2005

Yilan Liao,[1] Dongyue Li,[1,2] Ningxu Zhang,[1,2,3] Changfa Xia,[4] Rongshou Zheng,[4] Hongmei Zeng,[4] Siwei Zhang,[4] Jinfeng Wang[1] and Wanqing Chen[4]

Abstract

High-accuracy spatial distribution estimation is crucial for cancer prevention and control. Due to their complicated pathogenic factors, the distributions of many cancers' mortalities appear blocky, and spatial heterogeneity is common. However, most of the commonly used cancer mapping methods are based on spatial autocorrelation theory. Sandwich estimation is a new method based on spatial heterogeneity theory. A modified sandwich estimation method suitable for the estimation of cancer mortality distribution is proposed in this study. The variances of cancer mortality data are used to fuse sandwich estimation results from various auxiliary variables, the feasibility of which in estimating cancer mortality distributions is explained theoretically. The breast cancer (BC) mortality of the Chinese mainland in 2005 was taken as a case, and the accuracy of the modified sandwich estimation method was compared with that of the Hierarchical Bayesian (HB), the Co-Kriging (CK) and the Ordinary Kriging (OK) methods. The accuracy of the modified sandwich estimation method was better than the HB, the CK and the OK methods, and the estimation result from the modified sandwich estimation method was more likely to be acceptable. Therefore, this study represents an attempt to apply the sandwich estimation method to the estimation of cancer mortality distributions with strong spatial heterogeneity, which holds great potential for further application.

Keywords

Sandwich estimation, Bayesian, cancer mortality, spatial distribution, mapping

1 Introduction

Cancer has become a major global public health problem,[1,2] and its healthcare and financial burden will continue to increase in the coming decades.[1,3] In total, cancer caused 8.2 million deaths in 2012,[1] with the most commonly diagnosed cancers being lung cancer, accounting for 1.6 million deaths; liver cancer, accounting for 745,000 deaths; and stomach cancer, accounting for 723,000 deaths. Cancer is now the leading cause of death in China,[4] with both increasing incidence and increasing mortality. In general, cancer data

[1]The State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
[2]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China
[3]Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China
[4]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 authors:
Yilan Liao, The State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.
Email: liaoyl@lreis.ac.cn
Jinfeng Wang, The State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.
Email: wangjf@lreis.ac.cn

in China are collected by the National Central Cancer Registry of China (NCCR), and diagnoses are reported in administrative subdivisions from multiple sources, including local hospitals, community health centres, medical insurance providers and vital statistics.[4] However, the collected data are not geographically exhaustive, so accurate estimations of the spatial distribution of cancer mortality are important in understanding the complex pathogenic factors of cancer incidence and developing reasonable prevention and control policies.

Various methods of estimating the distribution of cancer mortality have been proposed,[5–8] including methods that only consider the spatial characteristics of cancer mortality surveillance data[6,9] and methods that consider the spatial association of relevant factors and cancer mortality.[7,8,10] Most methods only focus on spatial autocorrelation and rarely consider the heterogeneity of independent variables. However, spatial heterogeneity within samples is common in cancer distributions, especially over large areas of study.[11] Due to social economies, living habits and other factors, there are regional differences in the distribution of cancer mortality in China.[12,13] Therefore, methods that consider both spatial heterogeneity and various relevant factors of cancer mortality should be used to estimate the distribution of cancer mortality.

The sandwich estimation method, proposed by Wang,[14] is intended to be used on spatial stratified heterogeneous surfaces. Sandwich estimation can fully consider the spatial heterogeneity of cancer mortality by partitioning the study area into homogeneous subareas.[14,15] The sandwich estimation method can achieve high-precision estimation. Typically, one auxiliary variable or a couple variables cross-overlapping is sufficient for the sandwich estimation method. However, because of its complex geographical and pathogenic factors, more than 10 auxiliary variables strongly influence on cancer mortality. A single cross-overlap according to these auxiliary variables will result in over-zoning and is insufficient. Therefore, this study proposes a modified sandwich estimation method for estimating the distribution of cancer mortality.

Breast cancer (BC) is prevalent in China and is the most common malignant disease among women in developing countries.[2,16] The BC mortality rate is significantly spatially heterogeneous across China.[12,13] This paper considers spatial heterogeneity and the complex factors of cancer mortality, introducing a modified sandwich estimation method and using it to estimate the distribution of BC mortality in Chinese mainland in 2005. The ultimate purpose is to create an efficiency estimation method for mapping the distribution of cancer mortality in heterogeneous areas of study. At first, a set of factors (geographical factors, socioeconomic factors, physical conditions, living habits, etc.) were assumed to be the most likely to influence the distribution of cancer mortality. After a correlation analysis was applied to eliminate insignificant factors, several zoning layers were created and used in the sandwich estimation. The results were fused according to degrees of variance. A leave-one-out cross-validation (LOOCV) method was applied to validate the estimation method. The results showed that the modified sandwich estimation method could more fully consider spatial heterogeneity and the influences of factors on cancer mortality, performing better than the Hierarchical Bayesian (HB), the Co-Kriging (CK) and Ordinary Kriging (OK) methods.

2 Interpolation of cancer mortality

Estimating the distribution of cancer mortality is a type of disease mapping and, in practice, is a spatial interpolation problem. One of the most famous cases of disease mapping is John Snow's study of the nineteenth-century cholera epidemic in London,[17] but this study was not a full-coverage mapping. Spatial interpolation has been proposed as a solution to this problem. It predicts the values of unknown points or administrative units based on the known values of surveillance points or administrative units.[18] Many spatial interpolation methods have been used to estimate the distribution of cancer mortality. Although different interpolation methods have different basic principles, the common goal of these methods is representing the distribution of cancer mortality as accurately as possible.[19]

Some interpolation methods, such as inverse distance weighting (IDW) and the OK method, interpolate the distribution of cancer mortality based only on the spatial characteristics of surveillance data.[20,21] These methods adequately consider the spatial autocorrelation of cancer mortality surveillance data, which denotes the interdependence of data between locations, based on basic properties of geography.[11,22] Most spatial interpolation methods are based on spatial autocorrelation. However, spatial heterogeneity is another feature of geography,[11,23,24] especially for sampling across large areas.[11,14,25] Due to cancer's complex pathogenesis, regional differences are common in cancer mortality surveillance data. For example, there is a strong correlation between gastric cancer and diet,[26] and there is a significant difference in diet between southern and northern China; in addition, BC is strongly correlated with urbanization levels,[13,27] which vary greatly between urban and rural areas in China. Chien et al.[6] identified geographic variations in BC mortality across the USA.

While estimating the distribution of cancer mortality, some important pieces of information are easily overlooked such as spatial heterogeneity.

In addition, some spatial interpolation methods, such as linear regressions, estimate distributions with the help of auxiliary variables. Many auxiliary variables influence cancer mortality,[7,10,28] and a number of studies have used such variables to predict cancer mortality.[7,29,30] Further, many auxiliary variables data are more accessible than cancer mortality statistics.[28] Smith used a multivariable logistic regression to assess the correlations of age-adjusted bladder cancer mortality rates with socioeconomic, demographic and environmental variables.[29] The mortality rate of BC is also affected by various factors: physical condition,[31] genetic background,[32] eating habits,[33] living environment,[34] drinking,[35] smoking,[36] reproductive habits[37] and so on. Because of the increased availability of data, it is wise to estimate the distribution of cancer mortality by means of auxiliary variables. However, some linear regression methods consider the effects of auxiliary variables on cancer mortality independent of spatial correlation—a defect that should be remedied.

Of course, there are some methods that simultaneously consider spatial correlation and auxiliary variables, such as the HB method and the CK method. The HB method is recognized as one of the most powerful tools in disease mapping. It developed from the traditional Bayesian model,[38] which decomposes a complex estimation problem into a simple estimation problem that relies on conditional distributions of each parameter.[39] Spatial correlation was introduced into the HB method by defining spatial effects of cancer mortality in a study area. Specifically, in the HB method, the cancer mortality in any subarea relies on the cancer mortality in other subareas.[23] Liao et al.[40] estimated the distribution of neural tube defects in Heshun, China using the HB method, and Cross et al.[41] used it to analyze brucellosis in north-western Wyoming, but this method cannot completely eliminate random spatial noise. The homogeneity assumption of the HB method may not hold in a large study area, and a subarea in close proximity may be not the best subarea on which to rely.[14] The CK method uses auxiliary variables to improve the interpolation accuracy.[42] Knotters et al.[43] studied that the CK method using auxiliary variables would perform better than the OK mtheod. But the model will become very complex if there are too many auxiliary variables, and the relationship between target variable and auxiliary variables is vital to determine whether the CK method would be better than the OK method.[42,44] In addition, these methods may not apply to study areas in which the spatial heterogeneity is greater than the autocorrelation.

The sandwich estimation method, a new method suitable for spatial heterogeneous area, allows researchers to easily combine spatial heterogeneity and auxiliary variables,[14,15] and it has previously been used in disease mapping.[15] However, as too many auxiliary variables may lead to over-zoning, the sandwich estimation method still needs to be improved.

3 A modified sandwich estimation method

Estimating the distribution of cancer mortality often consists of two steps: collecting and analyzing data, and performing a suitable interpolation.[45–47] The main data that are used in distribution estimations of cancer mortality are cancer mortality surveillance data and auxiliary variables data. Cancer mortality surveillance data are collected by the NCCR, while data on auxiliary variables are taken from prior information or research literatures. The variables are then screened to identify useful variables and delete those that may negatively influence the accuracy of the estimation.[19] Performing the interpolation according to these data is the last step in obtaining the final result. The general framework for this study is shown in Figure 1, which is split into two parts: Variable selection and Sandwich estimation.

3.1 Variable selection

The pathogenesis of cancer is complicated, and geographical and pathogenic factors should be included in the estimation of cancer mortality distributions. However, the data for some variables are inaccessible, such as genetic background[32] and reproductive habits,[37] due to issues like privacy. Using prior knowledge and relevant studies, various accessible factors can be selected as auxiliary variables,[28,48] including geographical factors, socioeconomic factors, physical condition and living habits. As the quality of the auxiliary variables greatly influences the distribution estimation, it is necessary to select high-quality variables. Of course, high-quality early data collection processes are vital, but in many cases, the quality of the collected data is not controlled by the researcher. Therefore, the selection of variables becomes critical. The selection of variables primarily depends on the effects of the variables on cancer mortality. Two indicators are used to measure these effects: Pearson correlation coefficient and Geogdetector q-statistics.

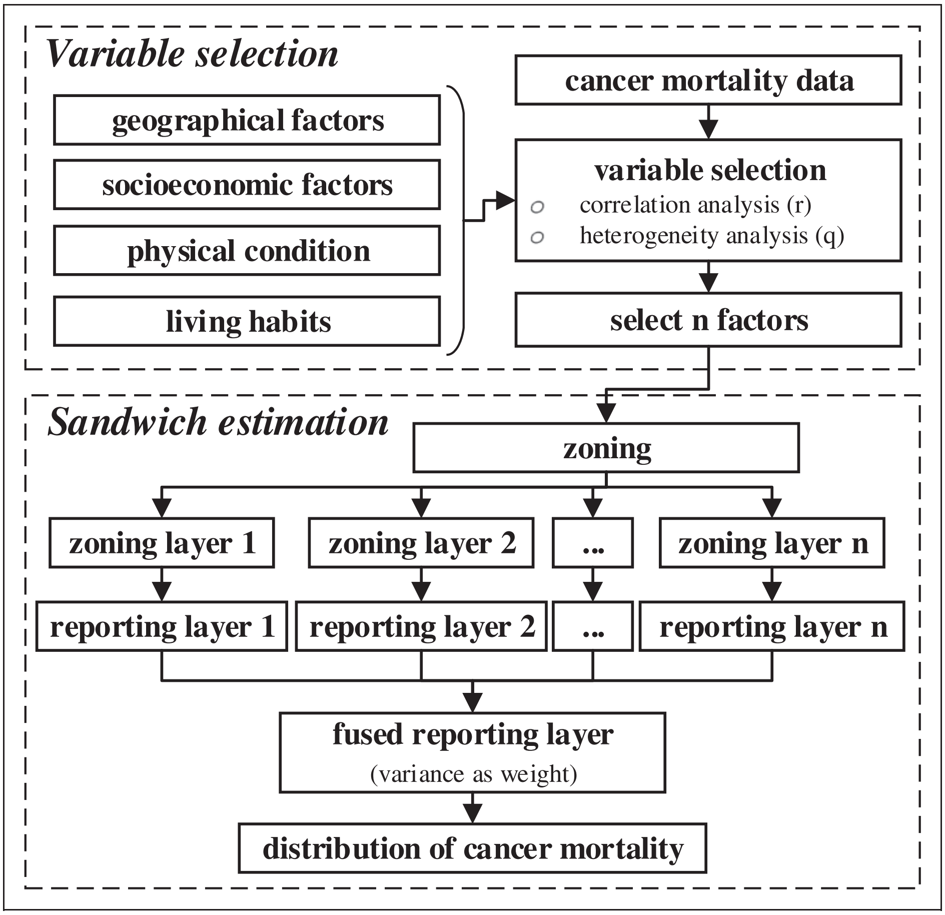

Figure 1. Sandwich estimation for estimating cancer mortality distribution.

3.1.1 Pearson correlation coefficient

The Pearson correlation coefficient (r) was developed by Karl Pearson in 1895 to measure the correlation between two datasets.[49] The Pearson correlation coefficient can be defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1}$$

where n is the magnitude of datasets X or Y, x_i and y_i indicate the i-st value of X and Y, respectively and $\overline{x}$ and $\overline{y}$ indicate the mean values of X and Y, respectively. The Pearson correlation coefficient is between -1 and 1. If $r > 0$, there is a positive correlation between the two datasets; if $r < 0$, there is a negative correlation between the two datasets; if $r = 0$, there is no correlation between the two datasets; if $r = \pm 1$, it means that there is a perfect correlation and a definite functional relationship between the two datasets. A larger absolute value of the Pearson correlation coefficient indicates a stronger correlation between the variable and cancer mortality, that is, the given variable has a greater effect on cancer mortality in some way.

3.1.2 Geogdetector q-statistics

Geogdetector q-statistics can be used to measure the spatial heterogeneity of cancer mortality and the influence of the geographical spatial partitioning of a given property on cancer mortality.[50,51] The q value quantifies the similarity between the spatial distribution of cancer mortality and a particular factor.[15,50] Geogdetector q-statistics can be defined as

$$q = 1 - \frac{\sum_{i=1}^{L} n_i {\sigma_i}^2}{n\sigma^2} \tag{2}$$

where n is the number of cancer mortality surveillance points in the study area, L is the number of zoning areas, n_i is the number of cancer mortality surveillance points in subarea i, σ^2 is the variance of cancer mortality surveillance points in the study area and $\sigma_i{}^2$ is the variance of cancer mortality surveillance points in subarea i. The value of q ranges from 0 to 1, and the closer q is to 1, the more similar the spatial distributions of cancer mortality and a particular factor are, which indicates that the factor has a greater influence on cancer mortality[51,52] and, at the same time, that the spatial heterogeneity of cancer mortality is stronger.

After preliminary variables are confirmed, auxiliary variables that have strong effects on cancer mortality can be screened out by combining r and q. The r and q are calculated using R 3.4.1 software.

3.2 Sandwich estimation

Sandwich estimation is a spatial estimation method for estimating stratified heterogeneous surfaces in multiple units. It comprises three layers, from which it gets its name[14]: a sampling layer, a zoning layer, and a reporting layer. The advantage of the sandwich estimation method is that multi-unit reporting can be achieved with few samples, providing a straightforward and simple way to solve the problem of multiple reporting units and the transfer of data between polygon systems.[14] The sampling layer is a collection of sampling points. The zoning layer divides the study area into multiple homogeneous zones of spatial attributes.[14,51] The values of the zoning layer can be estimated by taking the mean of the sampling data from the homogeneous layers. The reporting layer consists of spatial units,[14] which could be administrative units of a city, counties, ecological regions or watersheds. County administrative units were used as the reporting layer in this study. The sampling data is passed through the zoning layer to the reporting layer. The data and the error stream transfer the estimated mean and sampling error layer by layer.

3.2.1 Zoning layer

High-quality zoning is critical to achieving high-accuracy sandwich estimation.[14,53] Partitioning the study area into homogeneous subareas is the purpose of zoning,[53,54] which should consider the spatial structure of cancer mortality surveillance data. The zoning layer can be created using any zoning method, so long as it partitions the study area into homogeneous subareas.[14,15] There are three main tools of zoning[14,55]: prior information, pre-sampling and auxiliary variables. Prior information mainly stems from classical theories or previous research conclusions such as climate zone, and additional knowledge regarding partitions could be obtained by pre-sampling the units. After the selection of auxiliary variables believed to be correlated with cancer mortality, each auxiliary variable is partitioned into homogeneous subareas and combined with cancer mortality data.

In general, a zoning layer (according to one attribute variable or a cross-overlap of a couple of attribute variables) can meet the requirements of the sandwich estimation. Due to its complex geographical and pathogenic factors, more than 10 attributes strongly influence cancer mortality; a zoning layer with these attribute variables will lead to over-zoning and is insufficient for sandwich estimation. Therefore, several zoning layers are necessary to achieve high accuracy in the sandwich estimation. In this way, each variable corresponds to a zoning layer. For each zoning layer, the estimated mean $\overline{y_z}$ of zone z in this zoning layer is[14,15]

$$\overline{y_z} = \frac{1}{n_z}\sum_{i=1}^{n_z} y_{zi} \tag{3}$$

where n_z is the number of sampling points in zone z, y_{zi} is the sampling value at location i in zone z, and the estimated variance $V(\overline{y_z})$ of zone z is

$$V(\overline{y_z}) = E(\overline{y_z} - E\overline{y_z}) \tag{4}$$

For each auxiliary variable, one zoning layer estimation will be constructed from cancer mortality data.

3.2.2 Estimation layer

The reporting layer in the sandwich estimation method is flexible and could consist of administrative units, ecological grids, basins, and so on. The estimation layer in the modified sandwich estimation method is equivalent to the reporting layer in the standard sandwich estimation method. An estimation unit may overlap with one or more subareas from a zoning layer. The estimations for the estimation units (the administrative units in this study) were acquired using the zoning unit estimations.[14,15,56] The estimated value $\overline{y_r}$ of each unit in the estimation layer of one zoning layer can be expressed as

$$\overline{y_r} = \sum_{z=1}^{L_r} W_{rz}\overline{y_z} \tag{5}$$

$$V(\overline{y_r}) = \sum_{z=1}^{L_r} {W_{rz}}^2 V(\overline{y_z}) \tag{6}$$

where L_r is the number of zones that are covered or partially covered by estimation unit r in the estimation layer, $V(\overline{y_r})$ is the estimated variance of reporting unit r, and weight W_{rz} is

$$W_{rz} = \frac{A_{rz}}{\sum_{r=1}^{L_r} A_{rz}} \tag{7}$$

where A_{rz} is the area of zone z covered by the report unit r.

Although there is only one estimation layer, several zoning layers could create several sets of estimations for each unit in the estimation layer.

3.2.3 *Fused estimation layer*

Variance is a measure of the dispersion of a set of data.[57] In this study, variances specifically indicate the variations of cancer mortality data passing through various zoning layers in each estimation unit. Information, comprising estimations and sampling variances, flows from the sampling layer to the zoning layer and finally to the estimation layer.[14] Since the multiple zoning layers used in this paper can help to calculate estimations, variances flowed from each zoning layer according to each auxiliary variable and were used to compute the weights that fused the final estimations of each estimation unit. Therefore, the final estimated value y_r in each report unit r should be expressed as

$$y_r = \sum\nolimits_{j=1}^{n_v} W_{rj}\overline{y_{rj}} \tag{8}$$

where $\overline{y_{rj}}$ is the estimated value of report unit r from zoning layer j according to auxiliary variable j, n_v is the number of auxiliary variables and W_{rj} is the weight of $\overline{y_{rj}}$, which is defined as

$$W_{rj} = \frac{\frac{1}{V(\overline{y_{rj}})}}{\sum\nolimits_{j=1}^{n_v} \frac{1}{V(\overline{y_{rj}})}} \tag{9}$$

where $V(\overline{y_{rj}})$ is the estimated variance of report unit r from zoning layer j according to auxiliary variable j. In other words, a bigger variance indicates a smaller weight. In addition, the final variance $V(y_r)$ is calculated as

$$V(y_r) = \sum_{j=1}^{n_v} {W_{rj}}^2 V(\overline{y_{rj}}) \tag{10}$$

Further, in sandwich estimation, the sampling data is conducted by the zoning layer, and the zoning layer and estimation layer are mutually independent. Therefore, sampling data need not follow a specific spatial distribution. In the zoning layer, the distribution of sampling data within the partition also has no influence on the partition estimation. However, it is important to note that if a zone does not contain any sampling data, then the value for that zone cannot be estimated. The modified sandwich estimations are also performed using R 3.4.1 software.

4 Case study

4.1 Cancer mortality data

The estimation of the BC mortality rate distribution in the Chinese mainland in 2005 will serve as a case study to demonstrate the modified sandwich estimation. As a common and highly invasive malignant tumour with a slow progression of symptoms,[58] BC poses a serious threat to global women's health.[59] The morbidity of BC is low in China, but changes in eating habits, reproductive behavior and lifestyle are making people increasingly at-risk of developing BC.[60,61] Further, an accurate understanding of the distribution of BC mortality can provide a basis for research, prevention and treatment of BC.[28,62] However, due to differing economies, living habits and the complex pathogenic factors of BC,[63–66] there are regional differences in the distribution of BC mortality in China.[12,13]

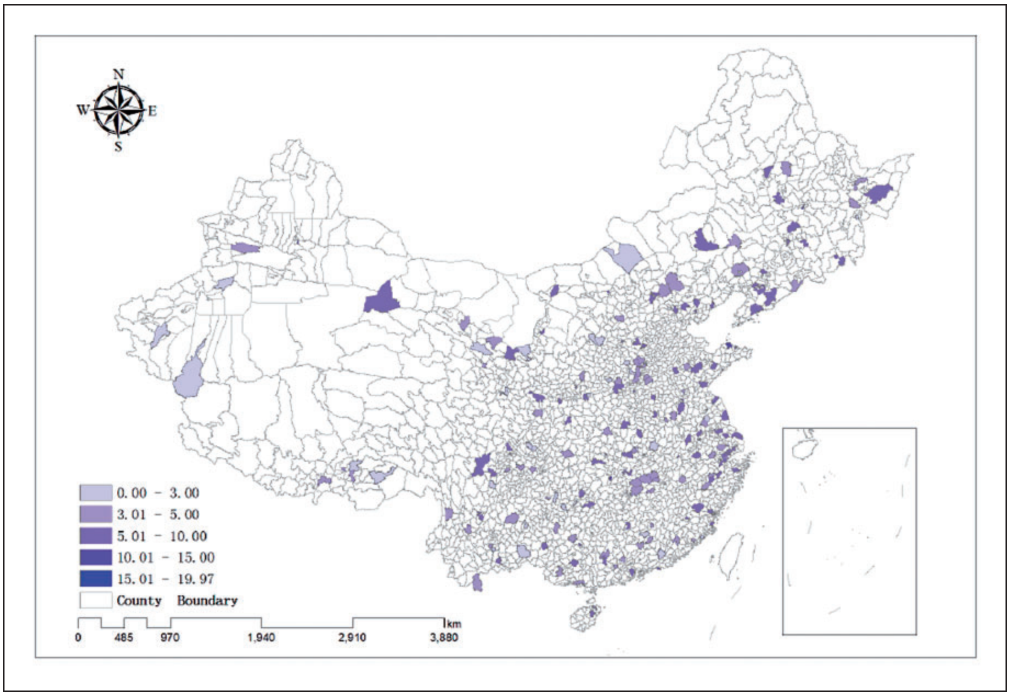

Figure 2. Distribution of BC surveillance counties in the Chinese mainland, 2005 (unit: 1/100,000).

The BC mortality surveillance data for 2005 was extracted from the Third National Retrospective Sampling Survey of Death Causes, conducted by the Chinese Ministry of Health from 2004 to 2005.[67,68] The BC mortality surveillance data covered 218 surveillance counties (Figure 2), accounting for 143 million person-years between 2004 and 2005.[67,68] The sampling survey collected information from a nationally representative sample.[68] As this is published tabulated data, no ethics approval was required for this study.[68] The distribution of the surveillance counties in the Chinese mainland is shown in Figure 2. Table 1 gives a description of the 2005 county-level surveillance BC mortality data from the Chinese mainland. The lowest county mortality of BC was 0.000 cases per 100,000, and the highest county mortality of BC was 19.975 cases per 100,000. And, the mean and median county mortality of BC were 5.658 and 5.135 per 100,000, respectively (Table 1), which indicated the data are not skewed. Spatial autocorrelation is common used in spatial populations[22] and can assist in the selection of interpolation methods.[18,69] In this study, Moran's I, which has a value between -1 and 1, is used to measure spatial autocorrelation.[22,70] The conceptualization of spatial relationship of Moran's I for BC mortality was inversed distance, which means that nearby neighboring features have a larger influence on the computations of a target feature than features that are far away. The threshold distance of Moran's I calculation was maximum of the nearest distance between all features. So, the data has a Moran's I of 0.213, which indicates a weak spatial autocorrelation; therefore, some interpolation methods based on spatial autocorrelation[71–73] may not able to accurately estimate the distribution of BC mortality. The Geogdetector q-statistic, which is used to measure spatial heterogeneity, has a value of 0.540, indicating great spatial heterogeneity in the distribution of cancer mortality. Hence, the modified sandwich estimation method may be suitable.

4.2 Auxiliary variables selection

BC has complicated pathogenic factors[63,64] that can generally be divided into two categories: 1) endogenous factors, such as genetic background,[32] physical condition[31] and hormone level,[74,75] and 2) exogenous factors, such as eating habits,[33] living environment,[34,65,66] drinking,[35] smoking,[36] reproductive habits[37] and other elements of lifestyle. However, many of these variables are inaccessible. Socioeconomic factors, who are positively related to BC risk, can serve as an auxiliary of lifestyle, which is generally inaccessible.[28,48] Genetic background encompasses a family history of cancer,[32] which is the primary cause of BC, but this is also difficult to access, primarily due to privacy issues.

Therefore, from the factors of physical condition, living habits, living environment and reproductive habits, 21 auxiliary variables were selected by reviewing the relevant literatures,[5,13,27,35,76–82] which are shown in Table 2.

Table 1. Summary statistics of BC mortality surveillance data ($n = 218$, unit: 1/100,000).

Cancer	Mean	Median	Min	Max	STD	Moran's I	q
BC	5.658	5.135	0.000	19.975	2.872	0.213*	0.540

*$P < 0.0001$.

Table 2. Pathogenic factors and corresponding auxiliary variables.

Pathogenic factor	Auxiliary variable	Abbreviation	Year
Physical condition	Female per capita body mass index (kg/m^2)	bmi	2004
	Female overweight rate (%)	over_weigh	2004
	Percentage of population aged 15–64 (%)	popul_15	2000
	Percentage of population over 60 years of age (%)	popul_60	2000
Living habits	Female smoking rate (%)	smoke	2004
	Female drinking rate (%)	Drink	2004
	Rate of excessive female red meat intake (%)	redmeat	2004
	Rate of insufficient female vegetable and fruit intake (%)	Fruit	2004
	Per capita milk intake (kilocalorie)	milk	2004
	Per capita animal fat intake (kilocalorie)	fat_intake	2004
Socioeconomic	Urbanization rate (%)	urban	2000
	Per capita gross national product (trillion dollars)	gdp	2005
	Average level of education for women (year)	edu	2000
	Proportion of non-agricultural population (%)	nonagri	2000
	Proportion of population in second industry (%)	sec_indus	2000
	Average population density (people/km^2)	popul_dens	2010
	Average number of live births of women aged 15–64	birth	2000
	Ratio of women over 15 years of age who have a spouse (%)	spouse	2000
	Proportion of dry land area (%)	dryland	2005
Geographical	Elevation (km)	elevation	2005
	Average annual fine particulate matter ratio (μg/m^3)	pm	2005

The data for these 21 auxiliary variables were taken from various sources. Socioeconomic data such as per-capita gross domestic product (GDP) were taken from the China Compendium of Statistics 1949–2008.[83] Demographic data, such as average number of live births for women aged 15 to 64, urbanization rates, the average level of education and female average fertility index, were extracted from the Fifth National Population Census (2000) and Sixth National Population Census (2010).[68] Health data such as smoking rates, alcohol consumption, fruit and vegetable intake and proportions of overweight women were extracted from the Nutrition and Health Status of the Chinese People (http://www.moh.gov.cn/wsb/pzcjd/200804/21290.shtml). The PM2.5 (particulate matter that has an aerodynamic diameter of 2.5 microns or smaller) data were extracted from the satellite inversion data of the National Aeronautics and Space Administration (NASA). The elevation dataset was provided by the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn). The corresponding used year for the auxiliary data were also shown in Table 2.

The Pearson correlation coefficient and the Geogdetector q-statistics of these auxiliary variables and BC mortality were calculated, and the results are shown in Table 3. It is worth noting that several auxiliary variables are continuous numerical data, which were first discretized using Ward's minimum variance method,[84–86] based on the theory that the sum of the squared distances in the class is the smallest. The results showed that 5–10 classes may be the best choice for zoning the BC mortality. For each variable, the zoning mode with the biggest Geogdetector q-statistics was selected as the zoning mode for that variable (Tables 3 and 4). Most of the auxiliary variables significantly influence BC mortality, with some exceptions, such as drinking, which seem to contradict common sense. According to the results, auxiliary variables (over_weigh, drink, fruit, milk, fat_intake and dryland) with insignificant correlations (p-value > 0.05) were removed from the study. Then, the auxiliary variable (redmeat) that was significantly correlated with BC mortality ($0.01 < p$-value < 0.05) but had only minor influence on it was also deleted. After selection, 14 auxiliary variables were retained. For these 14

Table 3. Pearson correlation coefficient and Geogdetector *q* statistics of auxiliary variables.

Auxiliary variable	*r*	*q*
bmi	−0.229[a]	0.262
over_weigh	0.042	0.267
popul_15	0.510[a]	0.442
popul_60	0.461[a]	0.324
smoke	0.193[a]	0.316
drink	−0.100	0.258
redmeat	−0.161[b]	0.286
fruit	−0.119	0.265
milk	−0.011	0.222
fat_intake	0.126	0.215
urban	0.546[a]	0.382
gdp	0.418[a]	0.330
edu	0.601[a]	0.451
nonagri	0.553[a]	0.393
sec_indus	0.435[a]	0.345
popul_dens	0.492[a]	0.490
birth	−0.643[a]	0.540
spouse	−0.242[a]	0.301
dryland	−0.072	0.308
elevation	−0.379[a]	0.345
pm	0.433[a]	0.378

[a]Correlation is significant at the 0.01 level (two-tailed).
[b]Correlation significant at the 0.05 level (two-tailed).

Table 4. Zoning layers for estimation.

Auxiliary variable	Zones	*q*
bmi	10	0.262
popul_15	10	0.442
popul_60	9	0.324
smoke	10	0.316
urban	10	0.382
gdp	10	0.330
edu	10	0.451
nonagri	9	0.393
sec_indus	10	0.345
popul_dens	10	0.490
birth	10	0.540
spouse	10	0.301
elevation	10	0.345
pm	10	0.378

variables, condition number (K) was calculated to determine the collinearity, generally, if $K < 100$, the degree of collinearity is small; if $100 \leq K \leq 1000$, there is a general degree of collinearity; if $K > 1000$, there is a serious collinearity. And the result showed that K was 45.8277, so there was very small collinearity between the auxiliary variables.

4.3 Sandwich estimation

From the three main sources of zoning,[14,55] this study chose to zone according to auxiliary variables. For the 14 auxiliary variables, 14 zoning layers were constructed. The BC mortality in each zone was roughly homogeneous to achieve the optimal partitioning mode of each variable. The partitioning mode of the auxiliary variables that had the greatest influence on BC mortality was found using Geogdetector q-statistics, with a greater q indicating a better fit. The optimal zoning situation of each auxiliary variable is shown in Table 4. As demonstrated in Table 4, most of the zoning layers (12 out of 14) were divided into 10 subareas and two zoning layers were divided into nine subareas.

The estimation layer consisted of 2862 counties. For each zoning layer, a set of values of the estimation layer would be estimated. As an estimation unit may overlap with one or more subareas from a zoning layer, the estimations were performed using equation (5). Several estimation layers were fused using equation (8).

4.4 Method comparison and precision evaluation

This study compares the effectiveness of applying the modified sandwich estimation method, the HB method and the OK method for estimating the distribution of BC mortality in the Chinese mainland in 2005. LOOCV was applied to evaluate the accuracy of the four methods, and the root-mean-squared-error ($RMSE$) and coefficient of determination (R^2) were calculated.

The HB method, recognized as one of the most powerful tools in disease mapping, is a statistical model written in multiple levels that estimates the parameters of the posterior distribution using the Bayesian method. This method defined probability distribution parameters of cancer mortality in the study area, with cancer mortality in any subarea relying on cancer mortality from other subareas.[23,39] Assumed that the BC mortality $O(i)$ in every region i is Poisson distributed[23,87]

$$O(i) \sim P(E(i) \cdot r(i)) \tag{11}$$

where $r(i)$ is the relative risk of cancer in region i, $E(i)$ is proportional to the total population of the region i, and $r(i)$ can be linearly transformed

$$\log(r(i)) = T + X^T U + v(i) + e(i) \tag{12}$$

where T is a constant value, the prior probability distribution of T is usually a flat distribution or a uniform distribution; X^T and U denote the auxiliary variables and corresponding coefficient, $v(i)$ reflects the spatial structure via an intrinsic Gaussian autoregression, and $e(i)$ reflects the heterogeneity of region i. Assuming that $w(i,j)$ is the spatial weight matrix that indicates the spatial adjacent relationship of location i and its n neighbors, if i and its neighbor j are spatial adjacent, then $w(i,j)$ is 1, otherwise is 0, $w^*(i,j) = w(i,j) \big/ \sum_{i=1}^{n} w(i,j)$ is the standardization of the rows of $w(i,j)$, so the prior distribution of $v(i)$ is[23]

$$v(i) \sim N\left(\sum\nolimits_{j=1}^{n} w^*(i,j)v(j), \kappa^2 \Big/ \sum\nolimits_{j=1}^{n} w(i,j)\right) \tag{13}$$

And the prior distribution of $e(i)$ is[23]

$$e(i) \sim N(0, \sigma^2) \tag{14}$$

where $1/\kappa^2 \sim Gamma(a,b)$, $1/\sigma^2 \sim Gamma(c,d)$, a and c are shape parameters, and b and d are inverse scale parameters.[23,88]

The OK method, based on spatial variation theory, is an optimized linear unbiased estimation method.[89,90] The OK method is one of the most common interpolation methods used to map the spatial distributions of attributes.[91–93] A condition of the OK method is that the regional variation $Z(x)$ be second-order stationary. If this condition is met, the OK method[94,95] can estimate BC mortality at any location using the weighted linear combination of observations from the study area

$$Z'(\mu_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{15}$$

where λ_i is the weight, which denotes the contributions from observations that ensure $Z'(\mu_0)$ is unbiased. The optimized unbiased estimation stands for the average of the estimated error or residuals close to zero, which is the mathematical expectation that the difference between the predicted value $Z'(\mu_0)$ and observed value $Z(x_0)$ is zero, and that the variation in the differences between the predicted $Z'(\mu_0)$ and observed $Z(x_0)$ is minimized.

The CK method is an extension of the OK method, also based on spatial variation theory.[89,90] These two methods are essentially the same. In addition to the target variable, the CK method is a multivariate kriging model that introduced several affected variables. If the regional variation $Z(x)$ is second-order stationary, the CK method can estimate BC mortality at any location using the weighted linear combination of observations from the study area

$$Z'(\mu_0) = \sum_{i=1}^{n} \lambda_{1i} Z_1(x_i) + \sum_{j=1}^{m} \lambda_{2j} Z_2(x_j) \tag{16}$$

where $Z_1(x_i)$ and $Z_2(x_j)$ are the measured value of target variable Z_1 and auxiliary variable Z_2, respectively; and λ_1 and λ_2 are the weight of target variable Z_1 and auxiliary variable Z_2, respectively. When the spatial information of a certain environmental variable that needs to be studied is missing, it can be analyzed by using the information of several variables related to it.[96] In theory, in a more complex large spatial scale area, the CK method will be superior to the OK method.[42,44]

The LOOCV, a method commonly used for evaluating the accuracy of interpolation methods, was applied in this study. A BC mortality surveillance point was removed from the original data; the remaining observations were used for the interpolations; and the value of the removed point was compared with corresponding estimated value.[97] The *RMSE*, one of the most commonly used evaluation methods, was calculated to compare the accuracy of predictions

$$RMSE = \sqrt{\frac{1}{n} \sum\nolimits_{i=1}^{n} (x_{obs,i} - x_{pred,i})^2} \tag{17}$$

where $x_{obs,i}$ denotes the observed BC mortality value in location i, $x_{pred,i}$ represents the estimated value in location i and n is the number of observations. A smaller *RMSE* indicates a more precise interpolation model. In addition, R^2 is often used in classical regression analyses, as it can measure the closeness of the model's fit to the variable. Here R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(x_{obs,i} - x_{obs,mean}\right)^2}{\sum_{i=1}^{n} \left(x_{obs,i} - x_{pred,i}\right)^2} \tag{18}$$

where $x_{obs,i}$ indicates the observed value in location i, $x_{pred,i}$ indicates the predicted value in location i, $x_{obs,mean}$ stands for the mean of the observed values and n is the number of observations. A larger R^2 indicates a more precise interpolation model.

4.5 Results and discussion

Four interpolation methods—the modified sandwich estimation, the HB, the CK and the OK methods—were used to estimate the distribution of BC mortality in the Chinese mainland in 2005. The modified sandwich estimation, the HB and the CK methods, used the same filtered auxiliary variables; the modified sandwich estimation method takes into consideration spatial heterogeneity; the HB method and CK method are based on the spatial autocorrelation, and the OK method only considers the spatial autocorrelation of the BC mortality surveillance data. Each method was calculated using R 3.4.1 software, the modified sandwich estimation method was performed using our own code, the HB method was performed using the R package-*INLA*, and the CK and OK methods were performed using the R package-*gstat*. The kriging and HB model parameters' settings are described in detail in the Supplemental Material.

The accuracy comparison of the four methods is shown in Table 5. The values of the *RMSE* of the four methods are 2.097 for the modified sandwich estimation, 2.116 for the HB method, 2.713 for the CK method and 2.779 for the OK method, indicating that the modified sandwich estimation method was the best method of them, then the HB, CK, and OK method. The R^2 ranked the methods in the same order as the *RMSE*. The linear regression graph of cross-validation (Figure 3) showed that the estimated BC mortalities from the modified sandwich estimation method and the OK method were relatively concentrated, while the results of the HB method and the CK method were relatively dispersed. The linear regression slope of the HB method was closest to 1, followed by the modified sandwich estimation method. The OK method had the smallest slope, indicating that the modified sandwich estimation method, the CK method and, particularly, the OK method

Table 5. Accuracy comparison of the four methods.

Method	*RMSE* (unit: 1/100,000)	R^2
Modified sandwich	2.097	0.587
HB	2.116	0.466
CK	2.713	0.145
OK	2.779	0.064

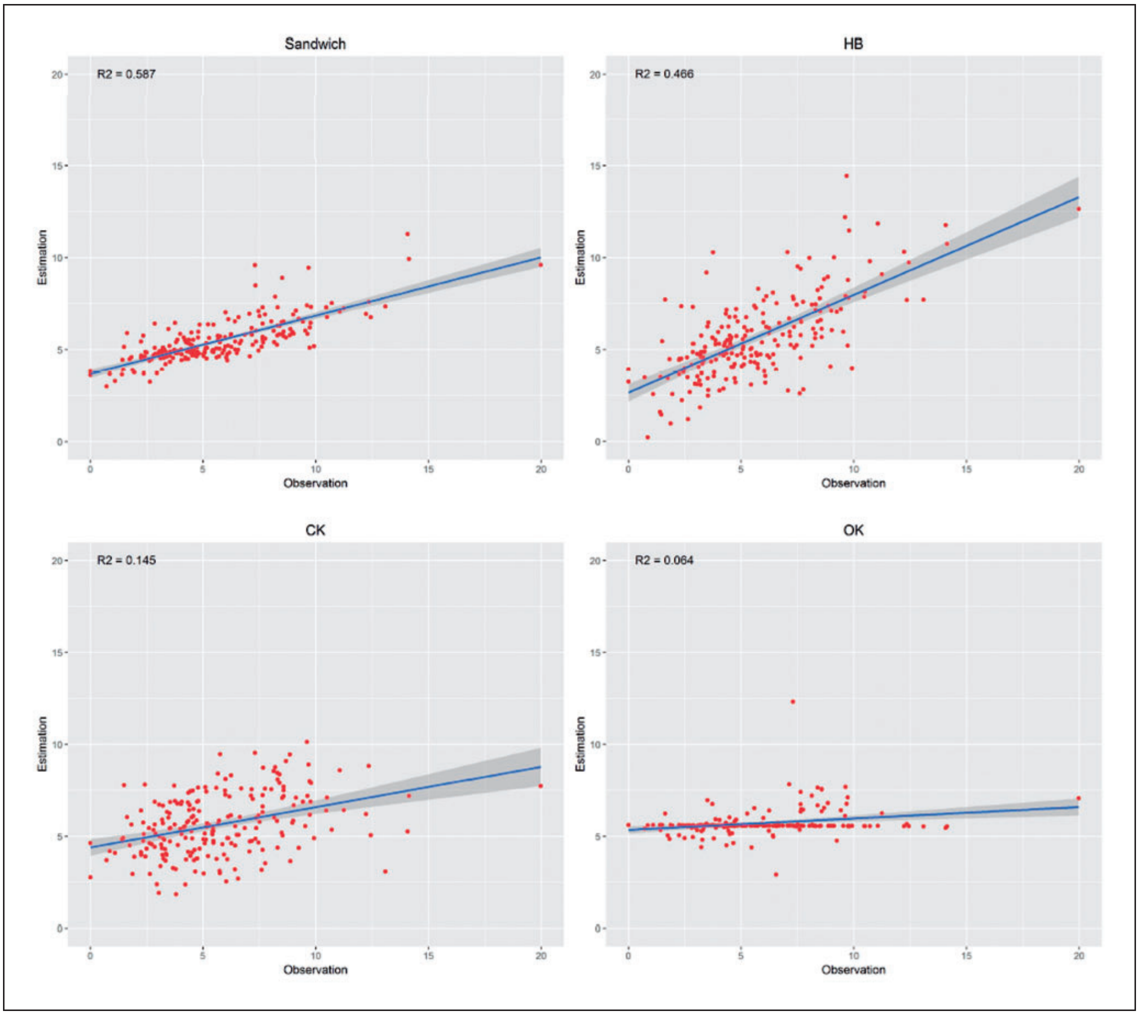

Figure 3. Linear regression graph of cross-validation (unit: 1/100,000).

underestimated BC mortality. However, in terms of precision, the modified sandwich estimation method is superior to the HB, the CK and the OK methods in estimating the distribution of BC mortality. More details about the precision of the four methods are included in the Supplementary Material.

The distribution of percentage differences between the estimated mortalities of the four methods and the surveillance data is illustrated in Table 6. In the surveillance data, the BC mortality of most counties (79.36%) was between 3 and 10 per 100,000. The estimated BC mortality of most counties (98.01% for the modified sandwich estimation method, 86.65%for the HB method, 92.17% for the CK method and 97.44% for the OK method) was between 3 and 10 per 100,000, but the estimated BC mortality of counties (7.37%) between 5 and 10 per 100,000 for the CK method was more than the other three methods, indicating the CK method underestimated the BC mortality seriously, and the estimated BC mortality of most counties (91.40%) was between 5 and 10 per

Table 6. Statistics of estimations from four methods.

Mortality (1/100,000)	Surveillance data	Modified sandwich	HB	CK	OK
0–3	33 (15.14%)	45 (1.57%)	325 (11.36%)	207 (7.23%)	35 (1.22%)
3–5	70 (32.11%)	1344 (46.96%)	1079 (37.70%)	976 (34.10%)	173 (6.04%)
5–10	103 (47.25%)	1461 (51.05%)	1401 (48.95%)	1662 (58.07%)	2616 (91.40%)
10–15	11 (5.05%)	11 (0.38%)	52 (1.82%)	16 (0.56%)	30 (1.05%)
>15	1 (0.46%)	1 (0.03%)	5 (0.17%)	1 (0.03%)	8 (0.28%)
Total counties	218 (100%)	2862 (100%)	2862 (100%)	2862 (100%)	2862 (100%)

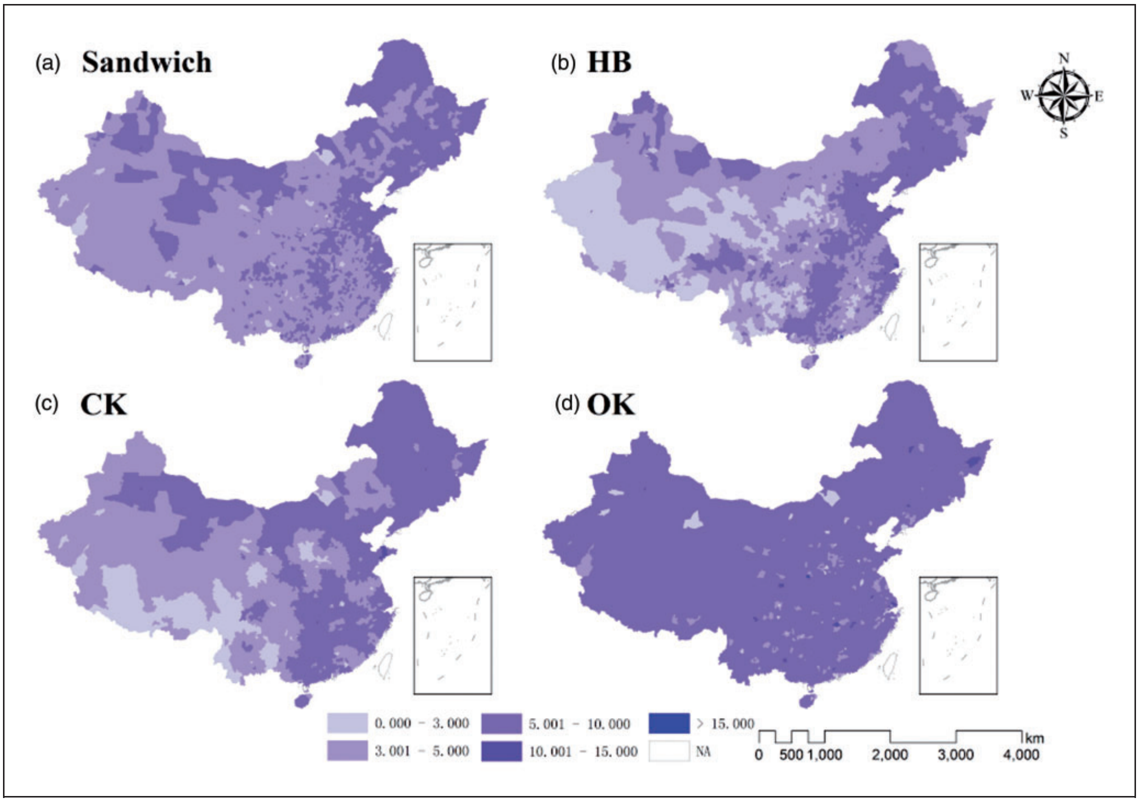

Figure 4. Distribution of the estimated BC mortality in the Chinese mainland in 2005 by (a) modified sandwich estimation, (b) HB, (c) CK and (d) OK (unit: 1/100,000).

100,000 for the OK method. This indicates that the results of the modified sandwich estimation method and HB method were closest to the BC mortality surveillance data. The four estimated distributions of BC mortality in the Chinese mainland in 2005 are shown in Figure 4. Comparing this figure with Figure 2, which shows the distribution of BC mortality surveillance data, can reflect the valuation of each method. Although the modified sandwich method underestimated the values in a few cases, it was more accurate than the other two methods.

According to the estimated BC mortality distribution of HB method (Figure 4(b)), the mortality of BC was low in some regions of central Inner Mongolia Province, western China, central and eastern China, while in some regions of north-eastern China, Inner Mongolia Province, North China Plain, Qinghai Province, Xinjiang Province, Chengdu Plain and south-eastern China, the BC mortality were relatively high. The HB method is able to describe complex interactions between the parameters of a stochastic model.[98,99] Compared to other methods, the HB method has two primary advantages[40,99]: first, it can manage the uncertainty of sampling data using a framework of probability theory; second, it relies on clear and definite assumptions. However, the HB method still needs to be improved. Ideally, the HB method would "leverage power" from similar but non-adjacent areas in a region. Like the OK method, the HB method also tends to smooth data.[100,101] The HB method only considers the spatial autocorrelation of adjacent regions, but in some situations this spatial autocorrelation does not exist,[23] which is the weakness of the HB method. When a study area contains heterogeneous characteristics, the HB method will inevitably enlarge the error of the smooth values. In this study, the accuracy of the HB method is worse than that of the modified sandwich estimation, but it still gets the same distribution as the result of the modified sandwich estimation method and its results are closest to the BC mortality surveillance data in cross-validation (Figure 3).

As previously mentioned, the CK and OK methods are primarily based on the theory of spatial autocorrelation, which allows them to make full use of the spatial autocorrelation of the BC mortality data.[102,103] Therefore, the ideal condition to utilize the CK and OK method is high spatial autocorrelation. The low spatial autocorrelation of the BC mortality surveillance data (Moran's I of 0.213) is not suitable for the CK and OK method, which led to the low precision for estimating the distribution of BC mortality (Figure 4(c) and (d) and Table 5). In addition, the CK method can also use the information of auxiliary variables, and it could perform better than the OK method (Table 5). Although the variation function of the kriging could be established with 218 BC mortality surveillance points,[104] this only encompasses 7.6% (218/2862) of the surveillance points, which is a relatively low sampling proportion. Meanwhile, the smoothing effect of the CK and OK methods is significant,[105] which means they can underestimate high sampling values and overestimate low sampling values,[106] as shown in Figure 4(c) and (d), especially the OK method.

Compared with the estimated BC mortality distribution of the HB, CK and OK methods, the BC mortality of the modified sandwich estimation method has a higher BC mortality interpolation values in the regions of eastern coastal areas of China, the north-eastern China, western Inner Mongolia Province, parts of Xinjiang Province and Gansu Province, and central China, which is consistent with some other researches[28,68,107] that have close or same study time. Although the spatial heterogeneity is common in geography,[14,23,70] it is often overlooked. The most commonly used interpolation methods are based on spatial autocorrelation theory. The hypothetical premise of the sandwich estimation method is that the heterogeneous areas can be zoned into several homogeneous subareas,[14] which is applicable to the mortalities of certain cancers.[6,13,26,27] Dirichlet is a stochastic process that is widely used for Bayesian nonparametric statistics.[108,109] It is a distribution over distributions,[108] which means that Dirichlet processes are often used to describe the prior knowledge about the distribution of parameters. Different from the Dirichlet processes, the sandwich estimation method can perform undeniably effective estimations on spatially heterogeneous surfaces.[14,15] The essence of the sandwich estimation method is calculating the mean of cancer mortality in the homogenous subarea. However, it is difficult to zone a study area into perfectly homogeneous subareas, which leads to a degree of smoothing (Figure 4(a)). In addition, due to over-zoning, an excess of zoning layers may not bring significant benefits to the sandwich estimation method.[110] This study improved the sandwich estimation method by solving the problem of over-zoning, increasing its application range and quality. Further, the comparisons to the HB and OK methods also show the advantage in accuracy of the modified sandwich estimation method. It is worth noting that sandwich sampling is preferred for the sandwich estimation method.[14,55]

5 Conclusion

Due to its complicated pathogenic factors, distributions of cancer mortality commonly exhibit spatial heterogeneity. Most of the traditional, commonly used interpolation methods, such as the HB, CK and OK methods, are based on spatial autocorrelation theory. Therefore, the spatial heterogeneity of cancer mortality may nullify the theoretical basis of these methods. The sandwich estimation method, based on spatial heterogeneity theory, transfers the cancer mortality surveillance data to the estimation layer through the zoning layer, which can be created simply and feasibly. However, in order to prevent over-zoning, large numbers of auxiliary variables should not be used in the original sandwich estimations, so it is sometimes difficult to utilize sufficient information in this method. The original sandwich estimation method was improved in this study: Each auxiliary variable now corresponds to a zoning layer; the corresponding estimation layers derive from the zoning layers; and finally, the reciprocals of the variances serve as the weights fusing these estimation layers and calculating estimated results. With BC mortality in Chinese mainland in 2005 as a case study, the accuracies of the HB, CK and OK methods were compared to that of the modified sandwich estimation method, which was found to be better than the HB method, and the HB method was found to be better than the CK and OK methods. In terms of distribution, the results from the HB and the modified sandwich estimation method are reliable and have almost the same distribution as the BC surveillance data. In short, this study attempted to use the sandwich estimation method to estimate the distribution of cancer mortality with strong spatial heterogeneity, which holds great potential for application in further estimations of cancer mortality distributions.

At the same time, there are some problems in this study. Using variances as weights may not be the best way to determine the weights used to fuse the results, and more effective weight determination methods should be explored in future work. The sandwich estimation method can eliminate some extreme values that should not be ignored. In addition, due to the limited data source, the auxiliary data whose date were close to data of 2005 were used, and the effect of BC incidence or death time lag should be included carefully in the future.

Authors' contribution

YL and DL contributed equally to this work.

Funding

The author(s) disclosed 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 (Project Number: 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 (Project Number: 2014FY121100).

ORCID iD

Dongyue Li http://orcid.org/0000-0001-8716-1381
Siwei Zhang http://orcid.org/0000-0002-7878-9940

Supplemental Material

Supplemental material for this article is available online.

References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015; **136**: E359–E86.
2. Jemal A, Bray F, Center MM, et al. Global cancer statistics. *Ca-Cancer J Clin* 2011; **61**: 69–90.
3. Bray F. Transition in human development and the global cancer burden. *World Cancer Report* 2014; 54–68.
4. Chen WQ, Zheng RS, Zhang SW, et al. Cancer incidence and mortality in China, 2013. *Cancer Lett* 2017; **401**: 63–71.
5. Martin-Sanchez JC, Cleries R, Lidon C, et al. Bayesian prediction of lung and breast cancer mortality among women in Spain (2014–2020). *Cancer Epidemiol* 2016; **43**: 22–29.
6. Chien LC, Yu HL and Schootman M. Efficient mapping and geographic disparities in breast cancer mortality at the county-level by race and age in the U.S. *Spat Spatio-temporal Epidemiol* 2013; **5**: 27.
7. Hegarty AC, Carsin AE and Comber H. Geographical analysis of cancer incidence in Ireland: a comparison of two Bayesian spatial models. *Cancer Epidemiol* 2010; **34**: 373–381.
8. Bristow RE, Chang J, Ziogas A, et al. Spatial analysis of advanced-stage ovarian cancer mortality in California. *Am J Obstet Gynecol* 2015; 213: 43.e1–43.e8.
9. Goovaerts P. Combining area-based and individual-level data in the geostatistical mapping of late-stage cancer incidence. *Spat Spatiotemporal Epidemiol* 2009; **1**: 61.
10. Eitan O, Yuval Barchana M, et al. Spatial analysis of air pollution and cancer incidence rates in Haifa Bay, Israel. *Sci Total Environ* 2010; **408**: 4429–4439.
11. Goodchild MF and Haining RP. GIS and spatial data analysis: converging perspectives. *Pap Reg Sci* 2004; **83**: 363–385.
12. Ying Zheng CW and Zhang M. The epidemic and characteristics of female breast cancer in China. *China Oncol* 2013; **23**: 561–569.
13. Fan L, Strasser-Weippl K, Li JJ, et al. Breast cancer in China. *Lancet Oncol* 2014; **15**: E279–E89.
14. Wang JF, Haining R, Liu TJ, et al. Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface. *Environ Plan A* 2013; **45**: 2515–2534.
15. Hu Y, Bergquist R, Lynn H, et al. Sandwich mapping of schistosomiasis risk in Anhui Province, China. *Geospat Health* 2015; **10**: 111–116.
16. Hong W and Dong ED. The past, present and future of breast cancer research in China. *Cancer Lett* 2014; **351**: 1–5.
17. McLeod KS. Our sense of Snow: the myth of John Snow in medical geography. *Soc Sci Med* 2000; **50**: 923–935.
18. Lam NSN. Spatial interpolation methods – a review. *Am Cartographer* 1983; **10**: 129–149.
19. Liao YL, Wang JF, Meng B, et al. Integration of GP and GA for mapping population distribution. *Int J Geograph Inform Sci* 2010; **24**: 47–67.

20. Xu MM, Guo YM, Zhang YJ, et al. Spatiotemporal analysis of particulate air pollution and ischemic heart disease mortality in Beijing, China. *Environ Health-Glob* 2014; 13: 1–12.
21. Bhunia GS, Kesari S, Chatterjee N, et al. Spatial and temporal variation and hotspot detection of kala-azar disease in Vaishali district (Bihar), IndiaBMC Infect Dis 2013; 13: 64.
22. Wang JF, Stein A, Gao BB, et al. A review of spatial sampling. *Spat Stat-Neth* 2012; **2**: 1–14.
23. Haining R and Zhang J. *Spatial data analysis: theory and practice*. Cambridge: Cambridge University Press, 2003, p.1077.
24. Wang JF, Zhang TL and Fu BJ. A measure of spatial stratified heterogeneity. *Ecol Indic* 2016; **67**: 250–256.
25. Foody GM. GIS: stressing the geographical. *Prog Phys Geog* 2004; **28**: 152–158.
26. Tsugane S and Sasazuki S. Diet and the risk of gastric cancer: review of epidemiological evidence. *Gastric Cancer* 2007; **10**: 75–83.
27. Fan L, Zheng Y, Yu KD, et al. Breast cancer in a transitional society over 18 years: trends and present status in Shanghai, China. *Breast Cancer Res Treat* 2009; **117**: 409–416.
28. Fei XF, Lou ZH, Christakos G, et al. A geographic analysis about the spatiotemporal pattern of breast cancer in Hangzhou from 2008 to 2012. *Plos One* 2016; 11: e0147866.
29. Smith ND, Prasad SM, Patel AR, et al. Bladder cancer mortality in the United States: a geographic and temporal analysis of socioeconomic and environmental factors. *J Urology* 2016; **195**: 290–296.
30. Pollan M, Ramis R, Aragones N, et al. Municipal distribution of breast cancer mortality among women in Spain. *BMC Cancer* 2007; 7: 1–14.
31. Cheraghi Z, Poorolajal J, Hashem T, et al. Effect of body mass index on breast cancer during premenopausal and postmenopausal periods: a meta-analysis. *Plos One* 2012; 7: e51446.
32. Lipkus IM, Iden D, Terrenoire J, et al. Relationships among breast cancer concern, risk perceptions, and interest in genetic testing for breast cancer susceptibility among African-American women with and without a family history of breast cancer. *Cancer Epidem Biomar* 1999; **8**: 533–539.
33. Christakos G and Lai JJ. A study of the breast cancer dynamics in North Carolina. *Soc Sci Med* 1997; **45**: 1503–1517.
34. Birnbaum LS. Developmental effects of dioxins and related endocrine disrupting chemicals. *Toxicol Lett* 1995; **82-83**: 743.
35. Pezzotti A, Kraft P, Hankinson SE, et al. The mitochondrial A10398G polymorphism, interaction with alcohol consumption, and breast cancer risk. *Plos One* 2009; 4: e5356.
36. Terry PD and Rohan TE. Cigarette smoking and the risk of breast cancer in women: a review of the literature. *Cancer Epidem Biomar* 2002; **11**: 953–971.
37. Yang RC, Mills PK and Dodge JL. Cancer screening, reproductive history, socioeconomic status, and anticipated cancer-related behavior among Hmong adults. *Asian Pacific J Cancer Prevent* 2006; **7**: 79.
38. Green PJ. Bayesian image-restoration, with 2 applications in spatial statistics – discussion. *Ann Inst Stat Math* 1991; **43**: 22–24.
39. Clark JS. Models for ecological data: and introduction. *Fish Fisheries* 2007; **9**: 328–329.
40. Liao YL, Wang JF, Wu JL, et al. A comparison of methods for spatial relative risk mapping of human neural tube defects. *Stoch Env Res Risk A* 2011; **25**: 99–106.
41. Cross PC, Heisey DM, Scurlock BM, et al. Mapping Brucellosis increases relative to elk density using hierarchical Bayesian models. *Plos ONE* 2010; 5: e10322.
42. Zhu Q and Lin HS. Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes. *Pedosphere* 2010; **20**: 594–606.
43. Knotters M, Brus DJ and Voshaar JHO. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. *Geoderma* 1995; **67**: 227–246.
44. Kravchenko AN and Robertson GP. Can topographical and yield data substantially improve total soil carbon mapping by regression kriging? *Agron J* 2007; **99**: 12–17.
45. Azpurua MA and Ramos KD. A comparison of spatial interpolation methods for estimation of average electromagnetic field magnitude. *Progr Electromagnet Res M* 2010; **14**: 135–145.
46. Li J and Heap AD. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. *Ecol Inform* 2011; **6**: 228–241.
47. Yao XL, Fu BJ, Lu YH, et al. Comparison of four spatial interpolation methods for estimating soil moisture in a complex terrain catchment. *Plos One* 2013; 8: e54660.
48. Hampton T. Studies address racial and geographic disparities in breast cancer treatment. *JAMA-J Am Med Assoc* 2008; **300**: 1641.
49. Rodgers JL and Nicewander WA. Thirteen ways to look at the correlation coefficient. *Am Stat* 1988; **42**: 59–66.
50. Wang JF and Hu Y. Environmental health risk detection with GeogDetector. *Environ Modell Softw* 2012; **33**: 114–115.
51. Wang JF, Li XH, Christakos G, et al. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *Int J Geograph Inform Sci* 2010; **24**: 107–127.
52. Ju H, Zhang Z, Zuo L, et al. Driving forces and their interactions of built-up land expansion based on the geographical detector – a case study of Beijing, China. *Int J Geograph Inform Sci* 2016; **30**: 2188–2207.
53. Li L, Wang J, Cao Z and Zhong E. An information-fusion method to identify pattern of spatial heterogeneity for improving the accuracy of estimation. *Stoch Env Res Risk A* 2008; **22**: 689–704.

54. Wang J, Wise S and Haining R. An integrated regionalization of earthquake, flood, and drought hazards in China. *Transact GIS* 1997; **2**: 25–44.
55. Wang JF, Haining R and Cao ZD. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. *Int J Geograph Inform Sci* 2010; **24**: 523–543.
56. Cochran WG. *Sampling techniques* (3d ed). New York, America: John Wiley & Sons, 1977.
57. Wolter KM. *Introduction to variance estimation*. New York, NY: Springer, 2007.
58. McPherson K, Steel CM and Dixon JM. ABC of breast disease: Breast cancer-epidemiology #risk |factors, and genetics. *Brit Med J* 2000; **321**: 624–628.
59. Ferlay JB, Bray P, Pizani P, et al. GLOBOCAN 2002: Cancer incidence, mortality and prevalence worldwide. Lyon, France: International Agency for Research on Cancer, 2004.
60. Zhang Q, Liu LY, Wang F, et al. The changes in female physical and childbearing characteristics in china and potential association with risk of breast cancer. *BMC Public Health* 2012; 12: 368.
61. Linos E, Spanos D, Rosner BA, et al. Effects of reproductive and demographic changes on breast cancer incidence in China: a modeling analysis. *J Natl Cancer Inst* 2008; **100**: 1352–1360.
62. Fei XF, Wu JP, Kong Z, et al. Urban–rural disparity of breast cancer and socioeconomic risk factors in China. *Plos ONE* 2015; 10: e0117572.
63. Adams J, White M and Forman D. Are there socioeconomic gradients in stage and grade of breast cancer at diagnosis? Cross sectional analysis of UK cancer registry data. *Brit Med J* 2004; **329**: 142–143.
64. Harper S, Lynch J, Meersman SC, et al. Trends in area-socioeconomic and race-ethnic disparities in breast cancer incidence stage at diagnosis, screening, mortality, and survival among women ages 50 years and over (1987–2005). *Cancer Epidem Biomar* 2009; **18**: 121–131.
65. Birnbaum LS and Fenton SE. Cancer and developmental exposure to endocrine disruptors. *Environ Health Perspect* 2003; **111**: 389.
66. Eriksen KT, Halkjaer J, Sorensen M, et al. Dietary cadmium intake and risk of breast, endometrial and ovarian cancer in Danish postmenopausal women: a prospective cohort study. *Plos One* 2014; 9: e100815.
67. Chen Z. *The third national retrospective sampling survey of death causes in China*. Beijing, China: China Union Medical University Press, 2008.
68. Xia C, Kahn C, Wang J, et al. Temporal trends in geographical variation in breast cancer mortality in China, 1973–2005: an analysis of nationwide surveys on cause of death. *Int J Env Res Pub Health* 2016; 13: 963.
69. Li J and Heap AD. Spatial interpolation methods applied in the environmental sciences: a review. *Environ Modell Softw* 2014; **53**: 173–189.
70. Daniel A, Griffith RH and Arbia A. Heterogeneity of attribute sampling error in spatial data sets. *Geograph Analys* 1994; **26**: 300–320.
71. Cattle JA, McBratney AB and Minasny B. Kriging method evaluation for assessing the spatial distribution of urban soil lead contamination. *J Environ Qual* 2002; **31**: 1576–1588.
72. Long W, Qin M, Jianjun Z, et al. Precise comparison of spatial interpolation for precipitation using KRIGING and TPS (thin plate smoothing spline) methods in Loess Plateau. *Sci Soil Water Conserv* 2011; **9**: 79–87.
73. Yasrebi J, Saffari M, Fathi H, et al. Evaluation and comparison of ordinary kriging and inverse distance weighting methods for prediction of spatial variability of some soil chemical parameters. *Res J Biol Sci* 2009; **4**: 93–102.
74. Whelan EA, Sandler DP, Root JL, et al. Menstrual cycle patterns and risk of breast cancer. *Am J Epidemiol* 1994; **140**: 1081–1090.
75. Bonnier P, Romain S, Dilhuydy JM, et al. Influence of pregnancy on the outcome of breast cancer: a case-control study. Societe Francaise de Senologie et de Pathologie Mammaire Study Group. *Int J Cancer* 1997; **72**: 720.
76. Barlow WE, White E, Ballard-Barbash R, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *J Natl Cancer Inst* 2006; **98**: 1204–1214.
77. Tian YF, Chu CH, Wu MH, et al. Anthropometric measures plasma adiponectin, and breast cancer risk. *Endocr-Relat Cancer* 2007; **14**: 669–677.
78. Montazeri A, Sadighi J, Farzadi F, et al. Weight, height, body mass index and risk of breast cancer in postmenopausal women: a case-control study. *BMC Cancer* 2008; 8: 278.
79. Suzuki R, Ye WM, Rylander-Rudqvist T, et al. Alcohol and postmenopausal breast cancer risk defined by estrogen and progesterone receptor status: a prospective cohort study. *J Natl Cancer Inst* 2005; **97**: 1601–1608.
80. Mathew A, Gajalakshmi V, Rajan B, et al. Anthropometric factors and breast cancer risk among urban and rural women in South India: a multicentric case-control study. *Brit J Cancer* 2008; **99**: 207–213.
81. Palmer JR, Adams-Campbell LL, Boggs DA, et al. A prospective study of body size and breast cancer in black women. *Cancer Epidem Biomar* 2007; **16**: 1795–1802.
82. Yuan JM, Wang QS, Ross RK, et al. Diet and breast-cancer in Shanghai and Tianjin, China. *Brit J Cancer* 1995; **71**: 1353–1358.
83. Statistics DoCSoNBo. *China compendium of statistics 1949–2008*. Beijing, China: China Statistics Press, 2010.
84. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 1963; **58**: 236–244.

85. Liao YL, Wang JF, Du W, et al. Using spatial analysis to understand the spatial heterogeneity of disability employment in China. *Transact GIS* 2017; **21**: 647–660.
86. Wang JF, Liu XH, Peng L, et al. Cities evolution tree and applications to predicting urban growth. *Popul Environ* 2012; **33**: 186–201.
87. Berke O. Exploratory disease mapping: kriging the spatial risk function from regional count data. *Int J Health Geogr* 2004; **3**: 18.
88. Hu Y, Wang JF, Zhu J, et al. Mapping under-five mortality in the Wenchuan earthquake using hierarchical Bayesian modeling. *Int J Environ Heal Res* 2011; **21**: 364–371.
89. Cressie N. Spatial prediction and ordinary kriging. *Math Geol* 1988; **20**: 405–421.
90. Oliver MA and Webster R. Kriging: a method of interpolation for geographical information systems. *Int J Geograph Inform Syst* 1990; **4**: 313–332.
91. Chen T, Liu XM, Zhu MZ, et al. Identification of trace element sources and associated risk assessment in vegetable soils of the urban-rural transitional area of Hangzhou, China. *Environ Pollut* 2008; **151**: 67–78.
92. Li WL, Xu BB, Song QJ, et al. The identification of ‘hotspots’ of heavy metal pollution in soil-rice systems at a regional scale in eastern China. *Sci Total Environ* 2014; **472**: 407–420.
93. Zhao KL, Liu XM, Xu JM, et al. Heavy metal contaminations in a soil-rice system: identification of spatial dependence in relation to soil properties of paddy fields. *J Hazard Mater* 2010; **181**: 778–787.
94. Delhomme JP. Kriging in the hydrosciences. *Adv Water Resour* 1978; **1**: 251–266.
95. Volpi G and Gambolati G. On the use of a main trend for the kriging technique in hydrology. *Adv Water Resour* 1978; **1**: 345–349.
96. Rossi RE, Mulla DJ, Journel AG, et al. Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecol Monogr* 1992; **62**: 277–314.
97. Mueller TG, Pusuluri NB, Mathias KK, et al. Map quality for ordinary kriging and inverse distance weighted interpolation. *Soil Sci Soc Am J* 2004; **68**: 2042–2047.
98. Gelman A, Carlin JB, Stern HS, et al. *Bayesian data analysis*, 2nd ed. London, Britain: Crc Pr I Llc, 2004.
99. Demichelis F, Magni P, Piergiorgi P, et al. A hierarchical Naive Bayes Model for handling sample heterogeneity in classification problems: an application to tissue microarrays. *BMC Bioinform* 2006; 7: 514.
100. Bernardinelli L and Montomoli C. Empirical Bayes versus fully Bayesian-analysis of geographical variation in disease risk. *Stat Med* 1992; **11**: 983–1007.
101. Clayton D and Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987; **43**: 671–681.
102. Griffith DA. *Advanced spatial statistics: special topics in the exploration of quantitative spatial data series*. Dordrecht, Netherlands: Springer Science & Business Media, 2012.
103. Robinson TP and Metternicht G. Testing the performance of spatial interpolation techniques for mapping soil properties. *Comput Electron Agri* 2006; **50**: 97–108.
104. Voltz M and Webster R. A comparison of kriging, cubic-splines and classification for predicting soil properties from sample information. *J Soil Sci* 1990; **41**: 473–490.
105. Goovaerts P. *Geostatistics for natural resources evaluation*. New York, America: Oxford University Press on Demand, 1997.
106. Lark RM and Webster R. Geostatistical mapping of geomorphic variables in the presence of trend. *Earth Surf Proc Land* 2006; **31**: 862–874.
107. Lai D. Geostatistical analysis of Chinese cancer mortality: variogram, kriging and beyond. *J Data Sci* 2004; **2**: 177–193.
108. Teh YW, Jordan MI, Beal MJ, et al. Hierarchical Dirichlet processes. *J Am Stat Assoc* 2006; **101**: 1566–1581.
109. Jara A. Theory and computations for the Dirichlet process and related models: an overview. *Int J Approx Reason* 2017; **81**: 128–146.
110. Li Shuhua HX, Qingbo Z, Bingbo G, et al. Nationwide soil moisture estimation based on Sandwich spatial estimation method in China. *Chinese J Agr Res Regional Paln* 2016; **37**: 8.