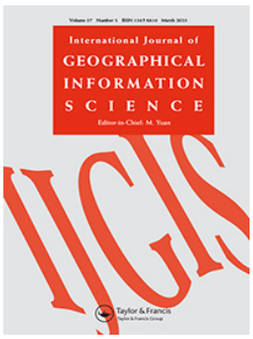

International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

A generalized heterogeneity model for spatial interpolation

Peng Luo, Yongze Song, Di Zhu, Junyi Cheng & Liqiu Meng

To cite this article: Peng Luo, Yongze Song, Di Zhu, Junyi Cheng & Liqiu Meng (2023) A generalized heterogeneity model for spatial interpolation, International Journal of Geographical Information Science, 37:3, 634-659, DOI: 10.1080/13658816.2022.2147530

To link to this article: https://doi.org/10.1080/13658816.2022.2147530

Published online: 20 Nov 2022.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2023, VOL. 37, NO. 3, 634–659
https://doi.org/10.1080/13658816.2022.2147530

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

A generalized heterogeneity model for spatial interpolation

Peng Luo[a], Yongze Song[b], Di Zhu[c], Junyi Cheng[d] and Liqiu Meng[a]

[a]Cartography and Visual Analytics, Technical University of Munich, Munich, Germany; [b]School of Design and the Built Environment, Curtin University, Perth, Australia; [c]Department of Geography, Environment and Society, University of Minnesota, Minneapolis, MN, USA; [d]Institute of Remote Sensing and Geographic Information Systems, Peking University, Beijing, China

ABSTRACT

Spatial heterogeneity refers to uneven distributions of geographical variables. Spatial interpolation methods that utilize spatial heterogeneity are sensitive to the way in which spatial heterogeneity is characterized. This study developed a Generalized Heterogeneity Model (GHM) for characterizing local and stratified heterogeneity within variables and to improve interpolation accuracy. GHM first divides a study area into multiple spatial strata according to the sample values and locations of a variable. Then, GHM estimates simultaneously the spatial variations of the variable within and between the spatial strata. Finally, GHM interpolates unbiased estimates and uncertainty at unsampled locations. We demonstrated the GHM by predicting the spatial distributions of marine chlorophyll in Townsville, Queensland, Australia. Results show that GHM improved both the overall interpolation accuracy across the study area and along strata boundaries compared with previous interpolation models. GHM also avoided bull's eye patterns and abrupt changes along strata boundaries. In future studies, GHM has the potential to be integrated with machine learning and advanced algorithms to improve spatial prediction accuracy for studies in broader fields.

ARTICLE HISTORY
Received 15 April 2022
Accepted 9 November 2022

KEYWORDS
Spatial interpolation; spatial heterogeneity; area-to-area kriging; stratified heterogeneity; spatial statistics

1. Introduction

Spatial prediction and interpolation play fundamental roles in geographic analysis (Lam 1983, Mitas and Mitasova 1999, Song 2022). An effective understanding of the characteristics of geographic variables guarantees the accuracy of the spatial interpolation (Oliver and Webster 1990, Zhu *et al.* 2020). Spatial dependence and spatial heterogeneity lay the foundation for spatial interpolations (Goodchild 2004, Tobler 2004). Geostatistical methods employ the spatial dependence of geographical variables for spatial prediction (Matheron 1963, Kyriakidis and Goodchild 2006). In kriging interpolation, widely used in geostatistics and comprising techniques such as ordinary kriging

CONTACT Yongze Song yongze.song@curtin.edu.au

(OK) and simple kriging, geographic variables are assumed to have second-order spatial stationarity for interpolation (Goovaerts 1997). Kriging assumes that the difference between the values of a geographical variable in two locations is independent of their locations but is only related to their distance (Goovaerts 1997). However, although the spatial second-order stationary assumption is typically satisfied in small areas, it may be weak in large areas with complex surfaces. Previous studies have developed methods to address the spatial non-homogeneity issue in interpolation tasks. For example, kriging with external drift (KED) removes spatial non-homogeneity through continuous drift (Hudson and Wackernagel 1994, Goovaerts 1997, Bourennane *et al.* 2000, Gao *et al.* 2020). However, the spatial stratified non-homogeneity is difficult to eliminate, owing to the continuous drift in the lower order (Chiles and Delfiner 2009).

Spatial non-homogeneity is often manifested by a geographic variable distributed over several spatial strata, each with homogeneous values. Geographical variables often show spatial stratification in the physical world, which is described as spatial stratified heterogeneity (SSH) (Wang *et al.* 2010). The characteristics of SSH make it difficult to construct a stable and reasonable semivariance function across the region (Gao *et al.* 2020). The spatial stratified strategy effectively predicts the spatial distribution at complex surfaces. SSH describes the geographical phenomenon that variable distributes as many homogeneous spatial strata with different spatial means or variances (Song *et al.* 2018, 2020b, Zhang *et al.* 2022). SSH does not require the assumption of spatial second-order stationarity in local heterogeneity. A few recent models have considered SSH to improve spatial interpolation. For example, in the stratified kriging (StK) algorithm, the entire study area is divided into several homogeneous strata, each of which is then subjected to interpolation (Liu *et al.* 2021). However, the numerical information of other strata is completely ignored when interpolating each stratum, and a stratum may only have a limited number of observations after the spatial partition. Ignoring the data between strata leads to limited information for constructing an accurate semivariogram and results in a loss of accuracy. In addition, the spatial division process and subsequent separate interpolation at each stratum may lead to unreasonably sharp changes along the strata boundaries (Gao *et al.* 2020).

Spatial dependence is still present in geographic factors located between the different strata, despite the existence of spatial stratification effect on a large scale (Song and Wu 2021). Geographical differences between strata are usually gradual, and stratification boundaries often manifest themselves as transition areas with certain widths (Fortin *et al.* 1996, De Smith *et al.* 2007, Hutchings *et al.* 2022). Geographic factors in transition areas usually have mixed characteristics with those of neighboring strata. This phenomenon is prevalent in both geographic and socio-economic factors. First, transition areas often exist between different strata of geographic factors, such as elevation and soil moisture. For example, elevation tends to decrease slowly from the plateau to the plain, creating a transition area. Second, spatial dependency between strata is also widely presented in socio-economic factors, such as land use (Preston 1966, Chen *et al.* 2020), nighttime light (Ma *et al.* 2015), economic development level (Erickson 1983) and population density (Luo *et al.* 2019). Significant differences exist between cities at different levels of development and between urban and rural areas (Hutchings *et al.* 2022). However, changes in the boundaries also tend to be gradual.

For example, population density and socio-economic levels tend to decline slowly from urban to rural areas, and there is a mixture of urban and rural characteristics at the urban-rural border. In summary, at larger scales, the distribution of geographical variables is spatially stratified, but there are usually continuous and gradual strata boundaries. However, current spatial interpolation models do not consider this phenomenon in geography.

The motivation of this study is to conduct accurate and reliable spatial prediction for large-scale geographic environments, considering both the existence of spatial stratification and spatial dependence at strata boundaries. A practical solution is to borrow information from other strata to consider both the spatial stratification strategy to ensure overall accuracy and reasonable estimates at the strata boundaries. Specifically, when performing spatial interpolation for strata boundaries, it is essential to consider information from different strata simultaneously. For example, when interpolating the population density in an urban-rural transition area, both urban and rural areas provide the necessary information. When interpolating elevations in the plains-plateau transition area, it is necessary to consider that the elevation in this area has a mixed characteristic of plateaus and plains.

With this motivation, two key issues need to be considered: the identification of transition areas or boundary areas, and the method used to borrow information from different areas. However, only a few studies have considered information borrowing to interpolate, and no study has considered the region in which borrowed information is needed. For example, a point mean of the surface with stratified non-homogeneity (P-MSN) algorithm was proposed to conduct interpolation in a large marine area (Gao *et al.* 2020). The study area was divided into several strata, and the semivariogram between each pair of strata was estimated using OK. It does not consider the existence of transition areas between partitions and assumes that the contribution of information from other regions to interpolation is offset by interference.

In summary, large-area stratified interpolation requires the process of bringing information from other strata, but this process often introduces high uncertainty in the result and leads to substantial computational cost. Therefore, trade-offs exist in the amount of information obtained from the outside stratum. The main concern is that observations from other strata or remote areas can introduce noise. We assume $n + k$ observations in the study area, including n observations in the interpolated stratum and k observations from other m strata. Previous studies have controlled the trade-offs by arranging different weights for the n observations within the stratum and k observations outside the stratum. This study provides a new method to automatically borrow information from other strata without manually adjusting the weights of different strata. The basic idea is to merge observations from each other strata separately and fit the semivariogram between two parts: the observations in the interpolated stratum and outside strata. Although all the observations from the outside strata are used to solve the spatial dependence between different strata, each stratum provides only one value in the fitted semivariogram. Thus, the uncertainty from the outside strata is expected to be limited when the information is borrowed. In addition, this approach reduces computational consumption and improves the interpolation efficiency.

Calculating the weights of other strata when conducting spatial interpolation was an important task in this study. Areal interpolation algorithms, such as area-to-point kriging (ATAP) and area-to-area kriging (ATAK) (Sadahiro 2000, Kyriakidis 2004, Goovaerts 2010), were developed to estimate the weights of areas. These algorithms were proposed to handle the interpolation of data at different scales and were used to disaggregate areal data into spatial prediction at the levels of points and different areas (Guan *et al.* 2011, Geddes *et al.* 2013, Hu and Huang 2020). Given its effectiveness in representing the spatial association between different areas, it is reasonable to believe that ATAK can be used to calculate the weights of other strata and characterize the spatial association between different strata.

In this study, a Generalized Heterogeneity Model (GHM) was developed. It combines ATAK and OK for the interpolation of spatial second-order non-homogeneity areas with high accuracy and efficiency. A specific geographical variable that presents spatial stratified non-homogeneity in a complex surface is distributed over many spatially homogeneous strata. Geographical variables that describe the same region exhibit spatial dependence, whereas variables that describe different regions exhibit spatial heterogeneity. The relationship between observations from different strata is represented by the relationship between strata. Thus, ATAK was introduced to characterize the spatial dependence between different strata and construct the corresponding semivariogram. In this way, information is borrowed while maintaining spatial dependence inside the homogeneous stratum. In addition, most of the information from other strata is noisy and interferes with interpolation accuracy. Using ATAK to characterize the spatial dependence between different strata may address this problem, because only the average value of each outside strata is considered in building the semivariogram.

We demonstrated the GHM using spatial interpolation of marine chlorophyll in Townsville, Queensland, Australia. Reliable and spatially continuous data on marine environments are essential for the conservation of biodiversity. However, in most marine areas, only sparse and unevenly distributed point samples are available, which is particularly pronounced in Australian marine regions (Li and Heap 2008). Therefore, it is critical to develop effective interpolation models for marine environments (Elumalai *et al.* 2017). Spatial interpolation in marine environments is challenging for two reasons. First, the spatial second-order stationary assumption is easily violated in large-area marine environments because of the highly dynamic movement of water masses and the resulting uneven distribution of ocean components (Gao *et al.* 2015, 2020). Stratification has been found and verified in marine environments (Bowman and Esaias 1981). Effective spatial interpolation technology that considers SSH is necessary for marine research. Second, spatial interpolation is a relatively difficult task in the marine environment, compared to that for the land environment, because of the lack of supporting explanatory variables. Without any supporting data, the mapping performance relies on understanding the characteristics of geographic variables through reasonable interpolation algorithms, which is an ideal case to verify the advantages of the proposed GHM.

The accuracy and effectiveness of GHM were evaluated through cross-validation and comparisons with previous related interpolation models, including OK, KED and

StK. The remainder of this paper is organized as follows. Section 2 describes the whole process of GHM for interpolation. Section 3 presents the implementation of GHM for the interpolation of marine chlorophyll in Townsville, Queensland, Australia. Section 4 discusses the findings and research contributions, and the study is concluded in Section 5.

2. Generalized heterogeneity model (GHM)

In this study, a Generalized Heterogeneity Model (GHM) was proposed to conduct stratified spatial prediction while considering information from other strata. This section is formulated as follows: concepts of GHM, development of the objective function, process of solving the function, optimal neighboring search strategy and execution of the GHM.

2.1. Concepts of GHM

Figure 1 shows the differences among classical geostatistical interpolation algorithms. The geographical data were assumed to be distributed as lower on the left and higher on the right. The interpolation theory of OK and KED is primarily based on spatial dependence, constructing semivariance functions at a global level. StK considers the existence of SSH by partitioning the space and constructing separate semivariance functions in each stratum to improve the accuracy. P-MSN considers that information

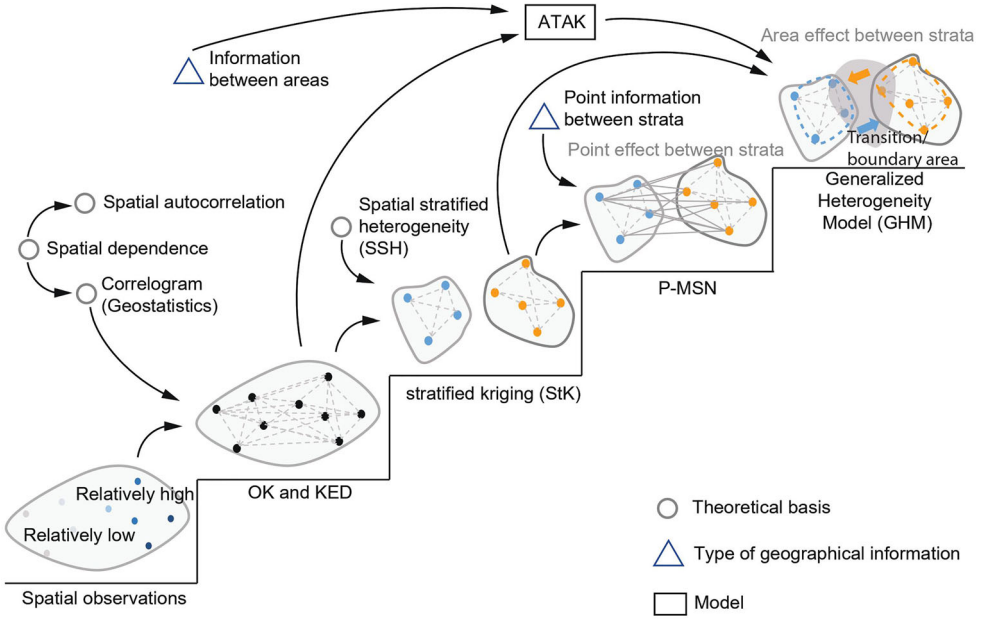

Figure 1. Theoretical basis of the Generalized Heterogeneity Model (GHM) and the relevant models: OK (ordinary kriging), KED (kriging with external drift), StK (stratified kriging), P-MSN (point mean of surface with stratified non-homogeneity) and ATAK (area-to-area kriging).

between different regions is borrowed from each other using OK to construct a semivariance function of the point level between strata.

GHM has two theoretical innovations compared to previous studies: (1) GHM considers the existence of strata boundaries (i.e. transition areas between strata) and the spatial dependence of strata at the boundaries. The borrowed information is used to improve the interpolation in these regions; (2) GHM borrows information from other strata in the form of an area. This has the promise of introducing valid information while avoiding interference information from other strata as much as possible.

2.2. Objective functions of GHM

Given that a spatially stratified area is divided into several homogeneous strata, the interpolated value is the weighted sum of two parts: observations within the interpolated stratum and observations outside the interpolated area. Assuming that spatial division has already been conducted, and there exist several homogeneous strata, the interpolated value is calculated as follows:

$$\hat{Z}_0 = Z_{in} + Z_{out} = \sum_{i=1}^{n} \lambda_i Z_i + \sum_{j=n+1}^{n+k} \lambda_j Z_j \tag{1}$$

where $\hat{Z}_0$ is the interpolated value, Z_{in} is the weighted sum of the observations in the interpolated stratum, and Z_{out} is the weighted sum of the observations in the other strata. n is the number of observations in the interpolated stratum, and k is the number of observations in the other strata. Z_i and Z_j are the observation values, where λ_i and λ_j are the weights of the observations.

Weight vector λ includes the weights of all observations, which are characterized as follows:

$$\lambda = [\lambda_{\text{in}}\ , \lambda_{\text{out}}\ ] = [\lambda_1, \lambda_2, \lambda_3 \ldots \lambda_n, \lambda_{n+1}, \ldots \lambda_{n+k}] \tag{2}$$

where λ_{in} is the weight vector of the observations in the interpolated stratum, consisting of λ_1,λ_2, $\lambda_3 \ldots \lambda_n$. λ_{out} is the weight vector of the observations in the other strata, consisting of $\lambda_{n+1}, \ldots \lambda_{n+k}$.

The interpolated values are estimated using the solved weight vector. Similar to other geostatistic models, two objective functions should be developed to obtain the best linear unbiased estimation:

$$\begin{cases} E(\hat{Z}_0 - Z_0) = 0 \\ \min\ \text{Var}(\hat{Z}_0 - Z_0) \end{cases} \tag{3}$$

By introducing the Lagrange multiplier, the two formulas are transformed into the following determinants (Appendices A and B):

$$\begin{bmatrix} R_{1,1} & \ldots & R_{1,n+k} & m_{s1} \\ R_{2,1} & \ldots & R_{2,n+k} & m_{s1} \\ \ldots & & & \cdots \\ R_{n+k,1} & \ldots & R_{n+k,n+k} & m_{s2} \\ m_{s1} & \ldots & m_{s2} & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \ldots \\ \lambda_{n+k} \\ L \end{bmatrix} = \begin{bmatrix} R_{1,0} \\ R_{2,0} \\ \ldots \\ R_{n+k,0} \\ m_{s1} \end{bmatrix} \tag{4}$$

where $R_{i,j}$ is the covariance between available observations i and $j(i,j = 1, \ldots, n+k)$.

$R_{i,0}$ is the covariance between interpolated point 0 and available observation $i(i = 1, \ldots n + k)$, where λ_i is the weight of the ith observation. m_{s1}, m_{s2} are the expectations of the variables inside and outside the interpolation stratum, respectively.

2.3. Solution

The matrix in the determinant (Equation (4)) describes three types of spatial dependence: dependence of observations in the interpolated stratum, dependence of observations between the interpolated stratum and the other strata and dependence of observations in the other strata. In geostatistical analysis, spatial dependence is described using a semivariogram.

We defined two kinds of semivariograms: the within-semivariogram S_w and the between-semivariogram S_b. S_w represents the spatial dependence of the observations in the interpolated stratum. S_b describes the spatial dependence between the observations in different strata, regardless of whether the observations in the interpolated stratum are included. In the equation for the determinant (Equation (4)), S_w includes the covariance of all the pairs of observations in the interpolated stratum. S_w includes the covariance between the two parts: observations in the interpolated stratum, and observations in the other strata. The S_w of each stratum was calculated using OK (Figure 2(c,d)).

S_b was solved by introducing ATAK. ATAK was initially used for interpolation using polygon data. To build a semivariogram between polygons, the polygons are disaggregated into points. The semivariogram between each pair of points is calculated and regarded as a semivariogram between polygons (Gotway and Young 2002, Yoo and Kyriakidis 2006). For example, in ATAK, the predictor of an area with an unknown value is calculated using a linear combination of covariances between nearby areas. The calculation of S_b in a stratum is the most important part of the GHM. First, all observations are merged into different areas (Figure 2(e,f)). Each observation in the

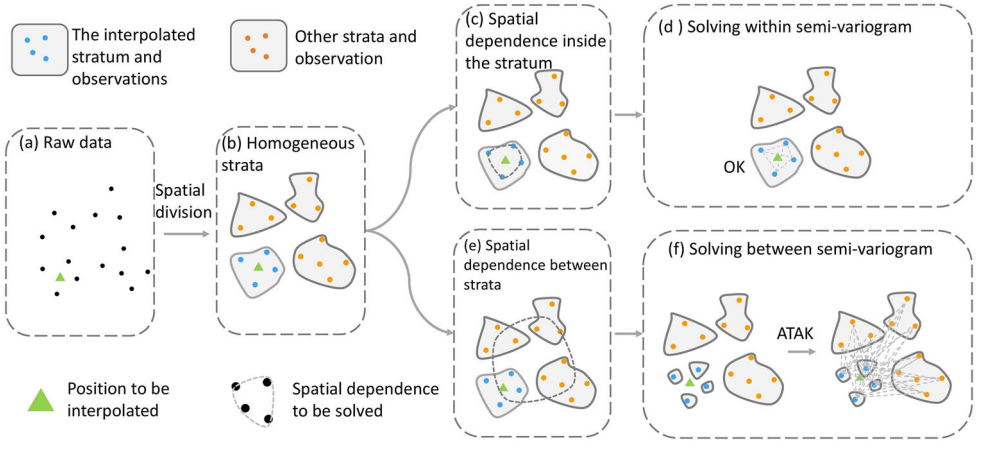

Figure 2. Semantic figure of the GHM for interpolation, including the spatial division, solving of the within-semivariogram S_w using OK, and the solving of the between-semivariogram S_b using ATAK.

interpolated stratum is regarded as an area with only one observation. Observations in other strata are separately merged into their respective strata. Second, the semivariance between the two areas is calculated using ATAK as follows:

$$R(\nu_i, \nu_j) = R(a_m, a_k) = \frac{1}{\sum_{s=1}^{P_m}\sum_{t=1}^{P_k}(w_s * w_t)} \sum_{s=1}^{P_m}\sum_{t=1}^{P_k}(w_s * w_t)R(u_s, u_t) \tag{5}$$

where ν_i and ν_j are the two observations, a_m is the stratum where ν_i occurs, and a_k is the stratum where ν_j occurs. If ν_i is the observation from the interpolated stratum, then it is equal to a_m. u_s and u_t are the observations of a_m and a_k, respectively; P_m and P_k are the numbers of observations of a_m and a_k, respectively and w_s and w_t are the weights of u_s and u_t, respectively, which are usually equal to one. u_s and u_t are necessary to estimate the area from discretized points.

It should be mentioned that although the observations from other strata are merged into several areas (i.e. strata), all observations are used to solve the spatial dependence between the different strata. Thus, the fit between semivariograms considers the spatial dependence in the interpolated stratum as much as possible and borrows information from the other strata.

2.4. Optimal neighboring search strategy

In geostatistical models, only the number of nearest observations (N_{max}) or observations within a certain range are used for interpolation, considering the computing efficiency (Lichtenstern 2013). As shown in Figure 3(a), only locations near the strata boundaries have neighboring observations from other strata and borrow information from other strata. In this study, we define an observation that may have neighboring observations from another stratum as a boundary observation, and if this condition does not hold, it is a non-boundary observation.

The N_{max} in the non-boundary area only controls information from the same stratum because only observations in the interpolated stratum are used for interpolation (Figure 3(a)):

$$\hat{Z}_{0_{nb}} = Z_{in} = \sum_{i=1}^{N_{\max}} \lambda_i Z_i \tag{6}$$

where $\hat{Z}_{0_{nb}}$ is the interpolated value in the non-boundary area.

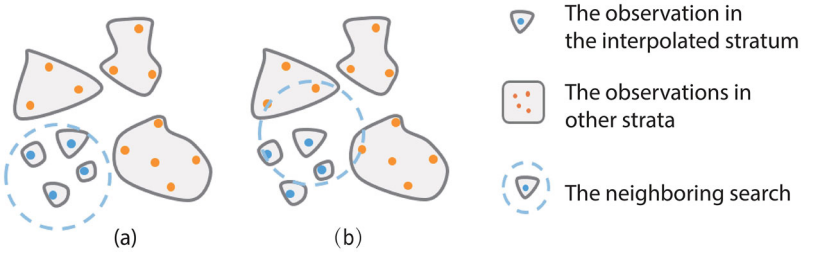

Figure 3. Influence of the neighboring search range to GHM: (a) observations that do not include borrowed information from other strata and (b) observations taking into account information from other strata.

In contrast, N_{max} in the boundary area determines how much information is borrowed from other strata (Figure 3(b)), because some neighboring observations are from other strata:

$$\hat{Z}_{0_b} = Z_{\text{in}} + Z_{out} = \sum_{i=1}^{n} \lambda_i Z_i + \sum_{j=n+1}^{N_{\max}} \lambda_j Z_j \tag{7}$$

where $\hat{Z}_{0_b}$ is the interpolated value in the boundary area.

Therefore, different search ranges (e.g., the number of nearest observations for interpolation) should be considered in the boundary and non-boundary areas. It is necessary to separately optimize N_{max} in the two areas. Optimal neighboring search strategy for boundary and non-boundary observations is proposed in this study. First, the boundary area of the interpolated stratum is identified (Figure 4(a)). For each stratum, the observations inside are divided into boundary area observations (Figure 4(a), dark color) and non-boundary area observations (Figure 4(a), light color). There are many methods for identifying boundary observations, eg edge detection for remote sensing images and buffer analysis of boundary lines. In addition, for sample point data, boundary identification is conducted depending on the number of neighboring observations from other strata or the distance to other strata. Second, after identifying the strata boundaries, the optimization of N_{max} in the boundary area (Figure 4(b)) and non-boundary area (Figure 4(c)) is executed. Different N_{max} values are set in the boundary and non-boundary regions; then, GHM interpolation is executed to obtain

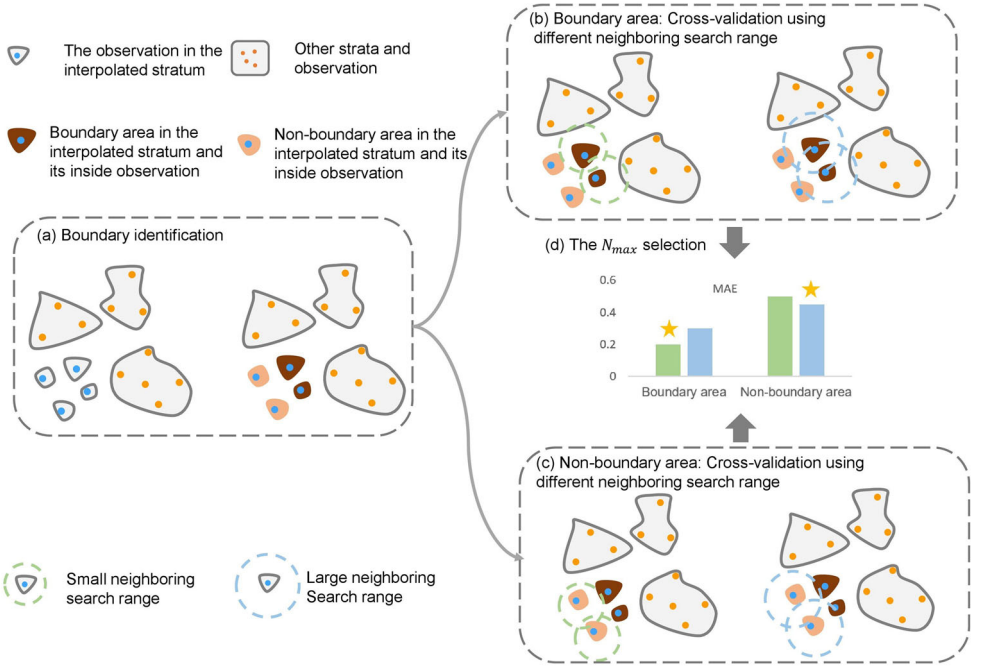

Figure 4. N_{max} selection process: (a) identify the boundary area of the interpolated stratum; cross-validation using different neighboring search ranges in (b) boundary areas and (c) non-boundary areas and (d) selection of N_{max}. The N_{max} with the highest validation accuracy is selected and labeled with a star.

the interpolation accuracy using cross-validation. Finally, the N_{max} values in the boundary and non-boundary areas with the highest accuracy are selected as the final N_{max} values for GHM interpolation (Figure 4(d)).

2.5. Execution process of GHM

The execution process of interpolation using GHM is summarized as follows. First, a large area, which is spatial second-order non-stationary, is divided into several homogeneous strata. The division process is conducted based on administrative units, geographical grids or expert experience and using clustering and image segmentation algorithms (Likas *et al.* 2003, Gao *et al.* 2020). Second, the semivariograms for each stratum are fitted. The variogram inside the stratum was fitted using OK. The variogram between different strata was fitted using ATAK.

Third, N_{max} optimization was conducted for each stratum. Finally, interpolation was conducted for each stratum. The interpolated value for the locations in the boundary area is the weighted sum of the neighboring observations inside and outside the stratum. The interpolated values in locations at non-boundary areas are the weighted sum values of the neighboring observations inside the interpolated stratum.

3. Case study: mapping marine chlorophyll using GHM

3.1. Study area and data

In this case study, we demonstrated GHM by spatial interpolation of marine chlorophyll in Townsville, Queensland, Australia. Marine chlorophyll data in the study area, including 4136 observations, were collected by the Australian National Facility for Ocean Gliders on 1 August 2010, which is a part of the Integrated Marine Observing System (IMOS) (Davies *et al.* 2018). The IMOS ocean observing mission is focused on the Australian coast and is critical for understanding the north-south transport of freshwater, heat and biogeochemical properties. These data are collected by sensors containing environmental information, such as temperature, chlorophyll, salinity and turbidity at different locations and instrument depths. The chlorophyll content ranges from 0.01 to 311.13, with an average value of 0.62, and the standard deviation is 5.12. Figure 5 shows the location of the study area and the spatial distribution of the marine chlorophyll observations in the study area. Figure 5 shows that a significant number of observations are located in very close proximity, considering that there are 4137 observations but only a few hundred that can be visually distinguished. Therefore, it is necessary to perform declustering prior to spatial modeling.

3.2. GHM-based interpolation

In this case, samples of marine chlorophyll observations are the only data used for spatial prediction. It is difficult to collect explanatory variables to support this prediction. Thus, the GHM provides an opportunity to accurately predict the spatial distribution of marine chlorophyll in the study area. The GHM-based interpolation of marine chlorophyll includes the following six steps (Figure 6): data pre-processing, spatial

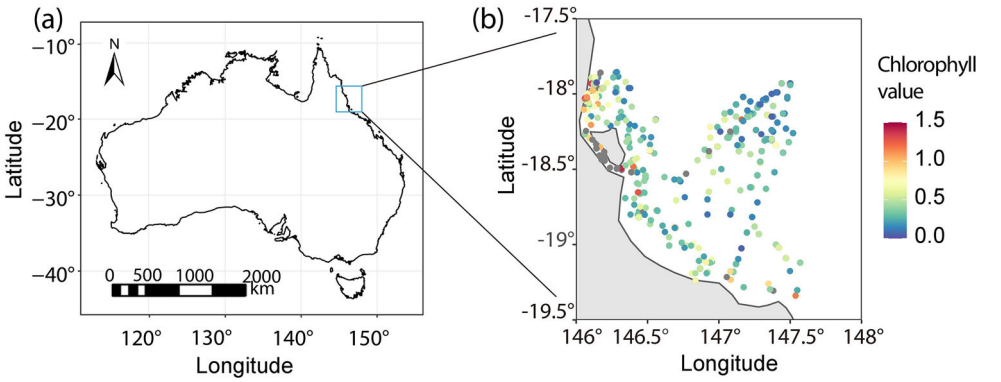

Figure 5. (a) Study area and (b) spatial distribution of marine chlorophyll samples in the study area. Observations with chlorophyll values out of the color legend (0 to 1.5) are shown as grey dots.

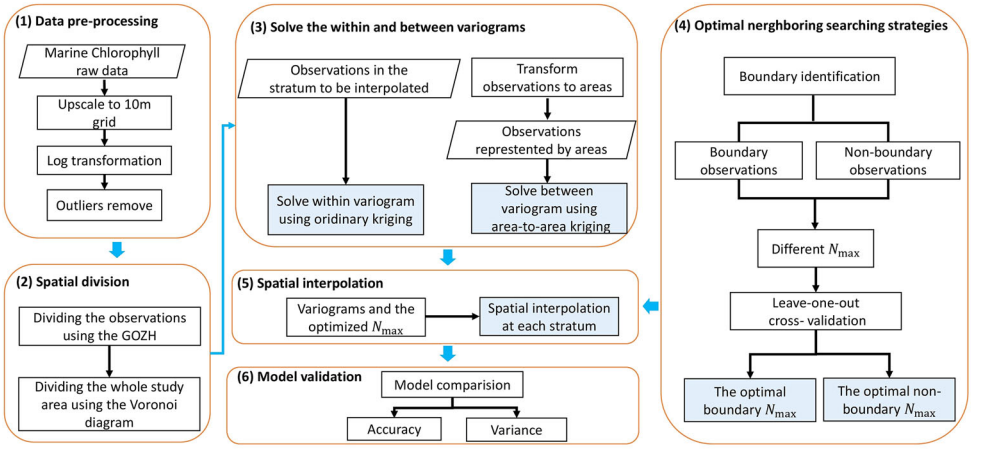

Figure 6. Flowchart of interpolation using GHM.

division, semivariogram solving, N_{max} optimization, spatial interpolation and accuracy assessment. These steps are introduced in the following paragraphs.

The first step is data processing. All observations were up-scaled to 10 m grids using an average composite. Then a log transformation was conducted because the original data approximately followed a log-normal distribution. Outliers were removed by eliminating any observations that were more than twice the standard deviation from the mean.

Second, a spatial division was conducted. An ideal partitioning result should have the smallest intra-partition variance and the largest inter-partition variance. Thus, a geographically optimal zones-based heterogeneity (GOZH) model was used to divide the entire area into several homogeneous strata (Wang *et al.* 2010, 2016, Song *et al.* 2020a, Luo *et al.* 2021, 2022). The GOZH model is a SSH model that allows spatial division that considers the maximum homogeneity within each stratum. Spatial

division is regarded as an optimization task in the GOZH model and is formulated as follows:

$$\Omega = \mathrm{Max}\left[1 - \frac{SSW}{SST}\right] \tag{8}$$

where Ω is a measure of the spatial stratified heterogeneity, SSW is the sum-of-squares within the stratum, and SST is the sum-of-squares total of marine chlorophyll in the whole study area. In the GOZH model, Ω was solved step-wise, with the same optimization objective, and was used to split the process of the Classification and Regression Tree method (Chipman *et al.* 1998, Luo *et al.* 2022). Spatial division was conducted after Ω was determined.

The spatial division guided by GOZH maintains spatial homogeneity inside each stratum as much as possible. Thus, the large spatial second-order non-homogeneous area was divided into several homogeneous strata. During the spatial division process, longitude and latitude were the two explanatory variables for marine chlorophyll. The entire study area was divided into several strata according to longitude and latitude using the GOZH model.

After the spatial division of the observations, a spatial division in the area without observations was conducted. In this study, we created a Voronoi diagram, using all the observations, in which the area of each diagram belongs to the same stratum as the corresponding observation. It should be noted that the GOZH-based spatial discretization method is not compulsive to be used in GHM. The optimal spatial discretization method should be selected according to the research question and corresponding expert knowledge. We chose the GOZH and Voronoi diagrams because they are intuitive and straightforward, obtaining the greatest non-homogeneity between different strata.

Third, the within- and between-semivariogram in each stratum was solved using OK and ATAK, respectively. For each stratum, the within-semivariogram S_w describes the spatial autocorrelation of all observations. The R package 'gstat' was used to build the semivariogram. The S_w varied with the strata. For a specific stratum, all observations inside were transformed into an area with a uniform value, and all other strata were merged into an area. Then, ATAK was used to construct a semivariogram between these areas. The R package 'atakrig' was used to conduct ATAK.

Fourth, the boundary observations were identified according to the number of nearest observations at other strata, and the optimal N_{max} was selected based on cross-validation. For a particular stratum, we counted the N nearest observations around each observation. The proportion of N observations from the other strata was then counted. We set a series of N values and chose 15 as the optimal value based on visual inspection, ensuring that the derived boundary area had a reasonable number of observations and a stable line structure. Hence, 15 neighboring observations were calculated for each observation, and the boundary observation was identified if at least one neighboring observation was from other strata. After identifying the boundary areas, the optimal N_{max} values for the boundary areas and non-boundary areas were identified. To select the optimal parameter, the range of N_{max} for the GHM is from 10 to 15. Because the largest N_{max} is smaller than 15, the identified non-boundary observation would have no neighboring observations from other strata. For each

stratum's boundary or non-boundary area, we set a different N_{max} and then performed GHM interpolation to verify and calculate the interpolation accuracy. Finally, we selected the N_{max} with the highest accuracy as the final interpolation parameter. In this study, leave-one-out cross-validation was used to optimize the parameters. Leave-one-out cross-validation is a particular case of k-fold cross-validation, in which the number of folds equals the number of observations (Wong 2015). Leave-one-out cross-validation is widely used to assess the interpolation performance of geostatistical models (Gong *et al.* 2014). Each observation was selected as the test set individually, and interpolation at this location was conducted using all other observations. In this study, the mean absolute error (MAE) derived from the leave-one-out cross-validation was used to compare the interpolation accuracy in the boundary and non-boundary areas to select the optimal N_{max}.

Fifth, interpolation was conducted within each stratum using the solved variograms and optimized N_{max}. Finally, the performance of GHM was evaluated using the leave-one-out cross-validation by comparison with three related geostatistical models, OK, KED and StK, which were conducted using the R package 'gstat'. In this study, StK shared the same spatial division result as GHM to fairly compare the performance. The semivariograms in OK and KED were solved using the R package 'gstat'. StK shared the same semivariograms at each stratum with the within-semivariogram of GHM. The N_{max} values for OK, KED and StK were selected according to sensitivity analysis. It should be mentioned that there was only one N_{max} in the entire study area for the three models, for both the boundary and non-boundary areas.

3.3. Results

3.3.1. Data pre-processing and neighboring search optimization

This section presents the results of data pre-processing, spatial division and N_{max} optimization. Figure 7(a,b) shows the process of spatial division using the GOZH model. The study area was divided into three strata, considering the highest homogeneity of marine chlorophyll within each stratum. The boundary identification results are shown in Figure 7(d). Most boundary observations were located in the regions from $-18\,\mathrm{N}^{\circ}$ to $-18.5\,\mathrm{N}^{\circ}$.

Figure 8 shows the process of N_{max} optimization for the three strata. In stratum A, the MAE in the non-boundary area was higher than that in the boundary area. The highest boundary MAE, at 0.635, corresponded to an N_{max} of 10. The boundary MAE decreased with an increase in N_{max} and reached its lowest value when N_{max} was 14. However, the MAE in the non-boundary area was very stable, ranging from 0.460 to 0.462. The lowest value of 0.460 was obtained when the N_{max} was 14. Compared with stratum A, there was no significant pattern of N_{max} in stratum B. The boundary MAE had the lowest value (0.435) when the N_{max} was 13. The non-boundary MAE had the lowest value (0.372) when the N_{max} was 11. In stratum C, the MAE in non-boundary was far higher than that in the boundary area, ranging from 0.516 to 0.529. Here, the lowest non-boundary MAE (0.516) corresponded to an N_{max} of 14. The boundary MAE increased with an increase in N_{max} from 10 to 15 and decreased slightly when N_{max} was 16. The lowest value, which is 0.268, corresponded to an N_{max} of 10.

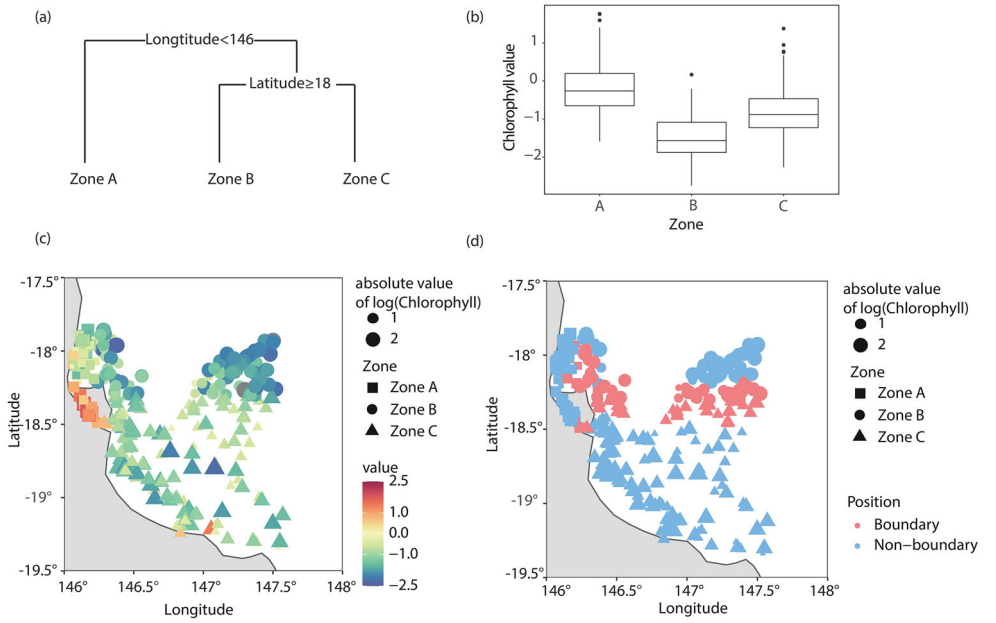

Figure 7. Data pre-processing and spatial division: (a) spatial division process based on the GOZH model; (b) box plots of marine chlorophyll data distribution in the divided strata; (c) observations belonging to three divided strata and (d) observations belonging to boundary and non-boundary areas.

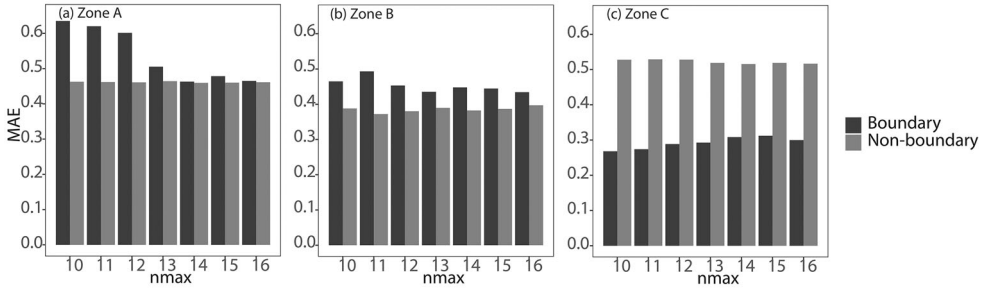

Figure 8. Selection of the optimal neighboring samples. N_{max} is the number of the nearest observations used for interpolation: (a) stratum A; (b) stratum B and (c) stratum C.

Table 1 lists the geostatistical parameters of the four models. The variogram of the original value was used by OK, and the variogram without drift was used by StK. For a specific stratum, the within-semivariogram S_w and between-semivariogram S_b characterize the spatial dependence within the interpolated stratum and between different strata, respectively. These two variograms were used by the GHM. In addition, the within-semivariogram S_w was also used by StK in each stratum.

3.3.2. Accuracy assessment and interpolation results

Table 2 shows the accuracy of the four models in the entire area, boundary area and non-boundary area. The two stratified models, StK and GHM, had better interpolation performance than the non-stratified models OK and KED. GHM had the highest

Table 1. Variogram of marine chlorophyll data under different conditions: variogram of the original value (OK), variogram without drift (StK), between-variogram (GHM) at each stratum, and within-variogram (GHM and StK) at each stratum.

Area	Type of variogram	Model	Sill	Nugget	Range (km)
Whole	Variogram of the original value	Sph	0.38	0.20	3.64
Area	Variogram without drift	Exp	0.35	0.00	3.30
	Within-variogram	Sph	0.27	0.00	0.66
Stratum A	Between-variogram	Sph	0.29	0.00	1.39
	Within-variogram	Sph	0.23	0.00	1.41
Stratum B	Between-variogram	Sph	0.30	0.00	12.95
	Within-variogram	Sph	0.32	0.00	0.95
Stratum C	Between-variogram	Sph	0.40	0.00	11.30

Table 2. Comparison of interpolation models, including OK, KED, StK and GHM, based on cross-validation.

	All area RMSE			Boundary area			Non-boundary		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
OK	0.485	0.626	0.392	0.415	0.563	0.440	0.501	0.640	0.376
KED	0.481	0.622	0.398	0.407	0.557	0.450	0.498	0.636	0.382
StK	0.465	0.600	0.424	0.380	0.509	0.531	0.485	0.620	0.397
GHM	0.457	0.589	0.439	0.374	0.505	0.536	0.476	0.607	0.414

accuracy among the four models, with the lowest MAE and root-mean-square (RMSE) values. The MAE values for the whole area, boundary area and non-boundary area were 0.457, 0.374 and 0.476, respectively. The MAE of the GHM for the entire area was 6.1%, 5.3% and 1.7% lower than OK, KED and StK, respectively. The RMSE of the GHM for the entire area was 6.3%, 5.6% and 1.9% lower than OK, KED and StK, respectively. In addition, GHM performed better interpolation in both the boundary area and non-boundary area. The MAE in the boundary and non-boundary areas for StK was 1.6% and 1.9% higher than that of GHM, respectively. GHM takes into account information from the other strata in the boundary, so the accuracy significantly increased, showing that marine chlorophyll in boundaries between strata has a spatial dependency, leading to smooth change. Borrowing information between different strata is necessary to improve the interpolation accuracy. The accuracy in non-boundary areas is increased owing to the use of the optimal parameter for the search area in the GHM. The accuracy of KED was slightly higher than that of OK in terms of lower RMSE and MAE. The MAE of KED was 0.83% lower than that for OK for the entire area. It performed better in the boundary area than in the non-boundary area, with an MAE 2.0% lower than that for OK.

Figure 9 shows the MAE in the boundary (Figure 9(a)) and non-boundary areas (Figure 9(b)) in the three strata. The GHM produces interpolation with the highest accuracy in all strata, especially in the boundary areas. The results showed that the stratified interpolation model significantly improved the accuracy of the boundary area. StK and GHM had lower MAE values than OK and KED. In the boundary area of stratum A, the interpolation MAE of GHM was 0.463, which was 8.4%, 12.5% and 13.0% lower than that of StK, KED and OK, respectively. In the boundary areas of strata B and C, the accuracies of GHM and StK were similar but were significantly higher

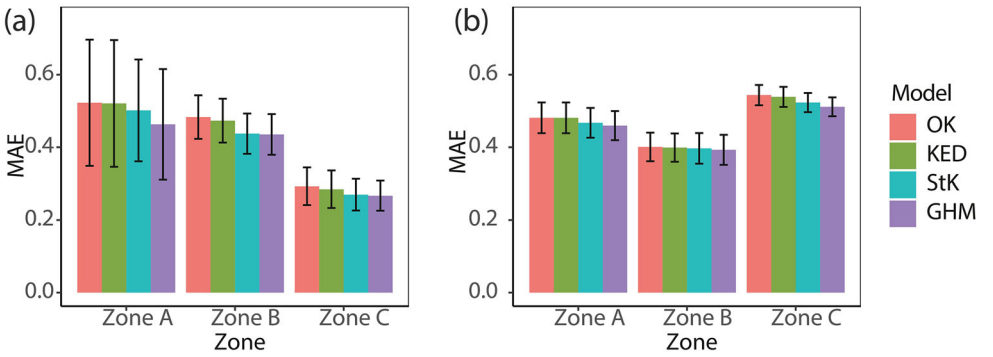

Figure 9. Comparison of the cross-validation MAE in different strata in the study area: (a) boundary area and (b) non-boundary area.

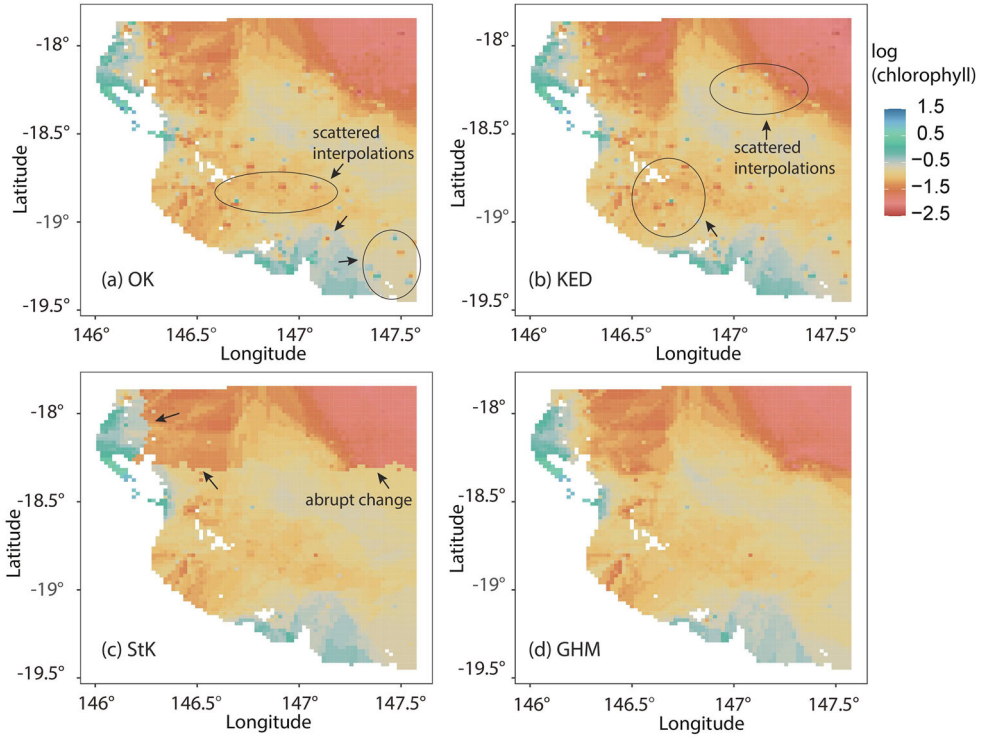

Figure 10. Spatial interpolation results of four models: (a) OK; (b) KED; (c) StK and (d) GHM.

than those of OK and KED. In the non-boundary area, MAE was still slightly lower than that of the other three models.

Figure 10 shows the interpolation results obtained by OK, KED, StK and GHM. Two non-stratified models, OK and KED, had smooth interpolation results because the study area was regarded as a whole, and only one semivariogram was built for the two models. However, the spatial prediction contained bull's eye patterns around the samples. The interpolation result from StK avoided the bull's eye patterns but showed abrupt changes along the boundaries between the strata. StK conducted the

interpolation in each region separately, and no information was borrowed from the other regions. Another stratified model, GHM, had a smooth result along the boundary, which was similar to the results of OK and KED. Ocean chlorophyll usually has a smooth distribution; therefore, the continuous change along the boundary is reasonable. In addition, the GHM avoided bull's eye patterns. In summary, the results demonstrate that our proposed GHM had the highest accuracy in both boundary area and non-boundary area and avoided bull's eye patterns and abrupt changes along the boundaries, enabling more reasonable spatial interpolations.

3.3.3. Interpolation uncertainty analysis

Figure 11 shows the spatial distribution of the estimation error from the GHM (Figure 11(a)) and the difference in absolute error between the GHM and the other three models (Figure 11(b–d)). As shown in Figure 11(b–d), GHM performed the best

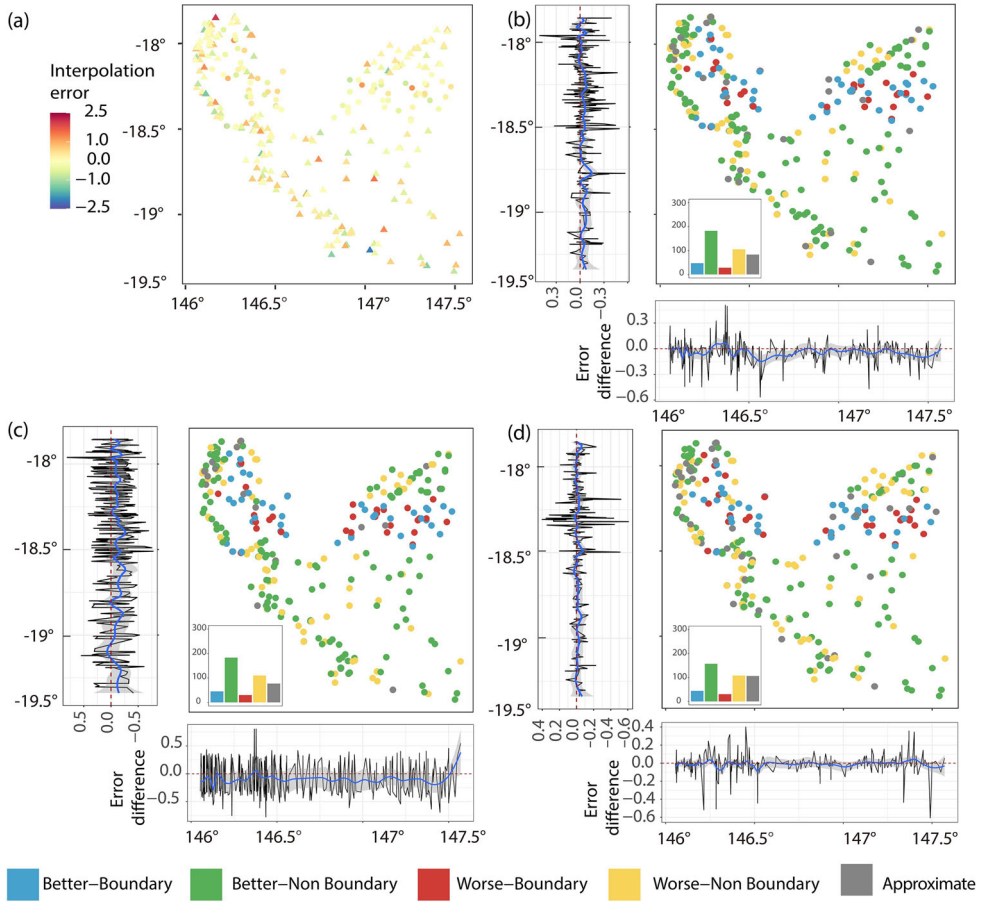

Figure 11. (a) Spatial distributions of the estimated errors of GHM, and the error difference between GHM and other models: (b) OK, (c) KED, (d) StK. An error difference lower than -0.05 means that GHM has better performance, and an error difference greater than 0.05 means the compared model has better performance. An absolute error difference within 0.05 shows that GHM has a performance similar to the compared model.

estimation (blue and green) among the four interpolation algorithms in most observations. Figure 11(b) shows a comparison of GHM and OK. GHM achieved better results at 51.5% of the observations than the other models. For the accuracy in the cross section, the estimation accuracy of GHM was higher than that of OK, except near 146.3 E°. The average accuracy was higher in all the regions. The division between regions A and B occurs near 146.3 E°. This indicates that the stratified process loses accuracy at the boundary between the two regions. From the longitudinal section, the average accuracy of GHM was higher than that of OK in most of the longitudinal cross sections. Figure 11(c) shows a comparison between GHM and KED. The average accuracy of the GHM was higher than that of the KED in most of the longitude and latitude cross sections.

Figure 11(d) shows the comparison of uncertainty for GHM and StK. Although the accuracy of GHM was higher than that of StK in the vast majority of the observed points, the absolute difference between the two estimation accuracies was not significant. The accuracy difference curves were around the value of zero in both the longitude and latitude cross sections. However, the uncertainty difference values were lower than zero in the majority of the regions, indicating that GHM had a relatively higher accuracy. It is worth mentioning that the accuracy of GHM was higher than that of StK at the boundary, between the latitudes -18 N° and -18.5 N°. This proves that the information-borrowing strategy of GHM was essential for reducing interpolation uncertainty.

The estimation variance is an essential indicator of interpolation performance. Figure 12(a) shows the variance in the observations derived from the GHM using

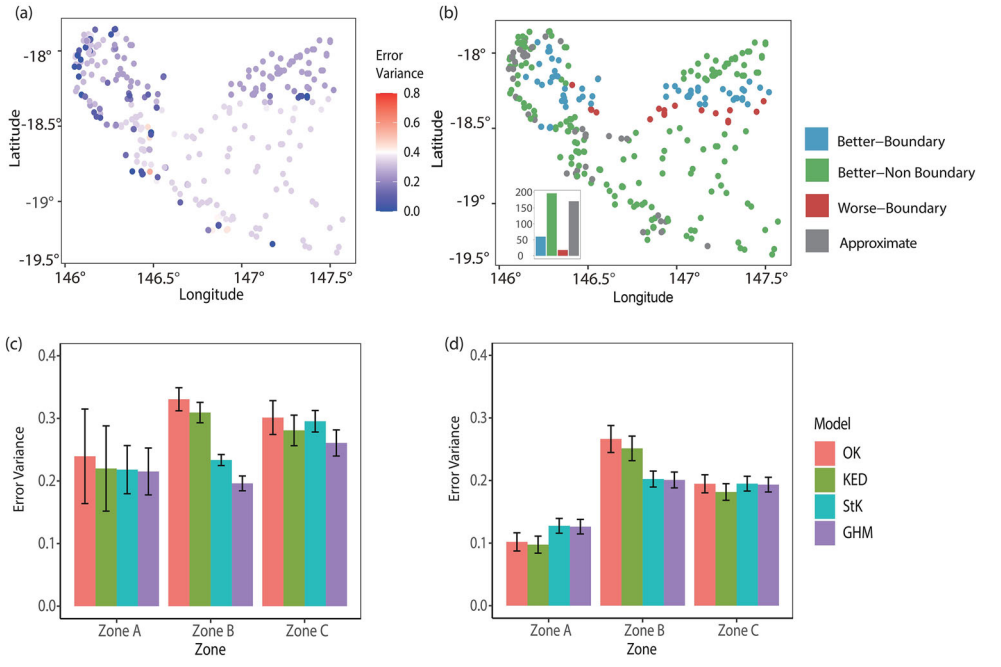

Figure 12. Cross-validation estimation variance of interpolations: (a) variance of GHM; (b) variance comparison between GHM and StK and variance in (c) boundary and (d) non-boundary areas in four models.

leave-one-out cross-validation. The variance ranged from 0 to 0.8. We compared the variance difference between the two stratified interpolation models, the proposed GHM and StK (Figure 12(b)). The results show that the GHM had a lower variance in most observations. GHM had a lower variance at 78% of boundary observations and 100% of non-boundary observations. A comparison of the error variance among the four models at different strata is shown in Figure 12(c,d). Generally, the two non-stratified models showed a higher average error variance than StK and GHM. Exceptions were the non-boundary areas of strata A and C. GHM showed the lowest average variance in most areas, including all boundary areas. The average variance of GHM (0.19) was significantly lower than that of the other models in the boundary area of stratum B, which was 42%, 39% and 17% lower than OK (0.33), KED (0.31) and StK (0.23), respectively. Two non-stratified models, OK and KED, had a lower variance in the non-boundary area of stratum A.

The estimation variance from the final interpolation process was mapped to the entire study area (Figure 13). Considering the estimation variance derived from leave-one-out cross-validation, the two stratified models have lower error variance than the non-stratified models. In the OK and KED models, the error variance at locations near observations was significantly lower than that at locations in other areas. In some areas with the highest estimation variance, such as the southwestern study area, GHM and StK also had a relatively low variance compared with OK and KED. GHM and StK had similar error variances in the non-boundary region because their interpolation

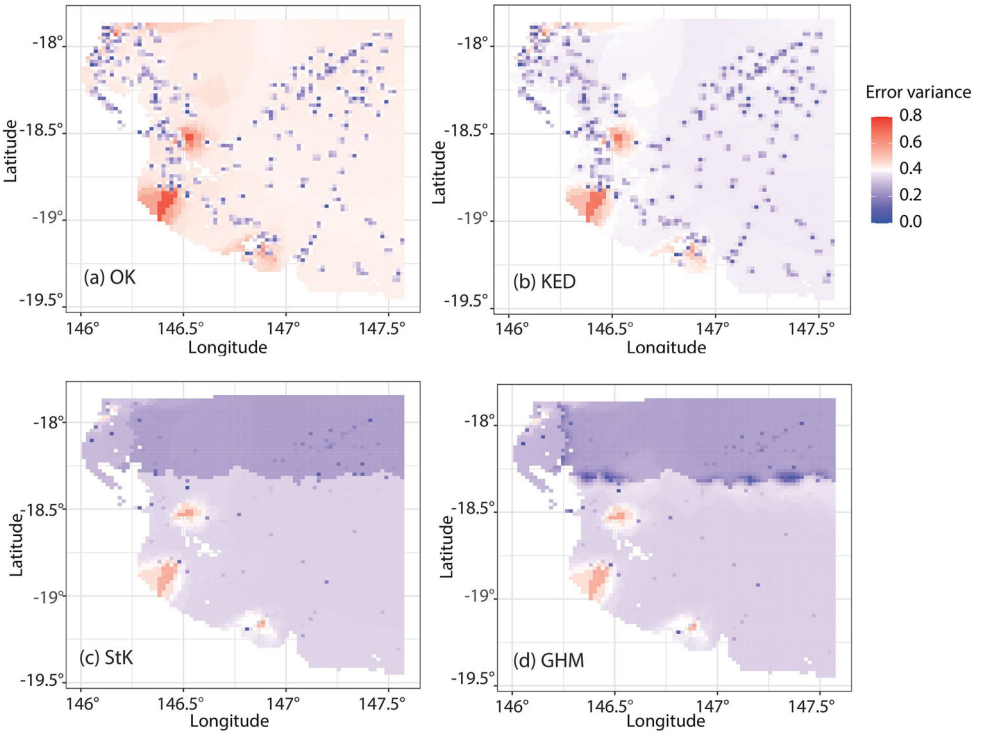

Figure 13. Spatial distributions of interpolation error variance derived from (a) OK, (b) KED, (c) StK and (d) GHM.

processes are quite similar apart from the boundary areas. GHM estimates interpolations with lower variance along the boundary than StK, because the boundaries are characterized and information from other strata is used in the GHM. However, in a remote area of the boundary region in stratum C, GHM had a higher interpolation variance than StK. The neighboring searching strategy arranges different N_{max} values to the boundary areas. The optimized N_{max} is effective for improving the overall estimation accuracy but may increase the uncertainty in the region close to the non-boundary area.

In summary, the results show that GHM has the highest estimation accuracy in terms of MAE and RMSE. GHM-based interpolations also had a lower variance, especially along the boundary regions. This demonstrates the effectiveness of the GHM and the necessity of borrowing information for the stratified geostatistical model.

4. Discussion

Spatial prediction is a challenging task for geostatistical models given that spatial second-order stationarity may be violated. Geographical variables tend to involve spatial stratification, with homogeneity within the stratification. Although heterogeneity exists between different strata, the stratification boundaries within geographical variables are bounded by spatial dependencies.

Previous studies have explored stratified interpolation algorithms, such as dividing the study area into homogeneous strata and removing continuous drift. However, the stratification process leads to information loss which limits the interpolation accuracy. Several methods have been developed to conduct the stratified interpolation while borrowing information from different strata. However, these methods ignore the spatial dependence that exists in the transition area between regions. In addition, when solving the kriging objective function, the constraints of these methods are typically too strong. In this study, we propose a GHM for interpolation in a spatially non-homogeneous large area. OK and ATAK were used to characterize the spatial dependence inside the interpolated stratum and between strata, respectively. The study area was divided into strata that were second-order stationary prior to interpolation. To interpolate each stratum, the semivariogram within the observations was solved using OK, and the semivariogram between observations from different strata was solved using ATAK. In addition, the boundaries between different strata were identified. The optimal neighboring observations (N_{max}) in the boundary and non-boundary areas was estimated using leave-one-out cross-validation.

In this study, we demonstrated the GHM through spatial prediction of marine chlorophyll in a study area in Australia. In similar cases, it is difficult to collect explanatory variables to support spatial prediction, and GHM performs well for spatial prediction. The results showed that the GHM had the highest interpolation accuracy in terms of RMSE and MAE. We found that the stratification strategy effectively improved interpolation accuracy in a large area with spatial second-order non-stationarity. Two stratified models, StK and our proposed GHM, had higher accuracies than OK and KED. They also had similar interpolation results in non-boundary areas. The GHM performed with a higher accuracy in the boundary area than StK. In addition, the interpolation

result from StK exhibited a sharp change along the boundary, resulting from spatial division. The GHM had a smoother estimation along the boundary because it borrowed information from other strata. A comparison of the error variances from the four models also verifies the necessity of information borrowing. The GHM had a lower estimation variance along the boundary than the StK, reducing the interpolation uncertainty. Apart from the three baseline models, we also compared the interpolation performance between the GHM and another information-borrowing model, the P-MSN. The results show that the MAE and RMSE of the P-MSN in the study area were 0.465 and 0.600, respectively. Its performance was similar to that of StK but lower than that of GHM. In addition, the MAE of GHM was 3.5% and 1.5% lower than that of P-MSN in the boundary and non-boundary areas, respectively. This indicates that the interpolation performance of the GHM is generally better than that of P-MSN, and the improvement is most evident in the boundary areas.

The main contributions of this study are summarized as follows: First, an effective and practical method for large-area mapping was developed by combining OK and ATAK. ATAK was used to characterize the spatial dependence between homogeneous strata. The introduction of ATAK is highly effective in large-area mapping. Second, an optimal neighboring search strategy was introduced to better borrow information from other strata when constructing the between-semivariogram S_b. Third, the results indicated that the influence of other homogeneous strata might be characterized as an area effect.

Spatial second-order stationarity is challenging in large areas, and a single semivariogram cannot reflect the real spatial dependence of the entire area. Dividing the region into several homogeneous small regions is a straightforward way to improve interpolation accuracy, which is the basic idea of StK. However, StK may lead to each stratum having limited observations, which introduces the difficulty of fitting a semivariogram. In addition, conducting interpolation at different strata leads to a sharp change along the stratum boundary. In some special environments, such as cliffs and faults, environmental variables may exhibit sharp changes. In such a situation, spatial division is conducted according to special geography, and the sharp change in interpolation results from StK might be reasonable. However, most geographical variables change gradually in the real world, especially marine environmental variables such as the chlorophyll selected in our case study. This sharp change is unreasonable for the spatial distribution of most geographical variables. Borrowing information from other strata along the boundary is the basic idea of the GHM, and its effectiveness was verified by our results.

However, there are still some limitations to this study. First, the methods used for the spatial division and optimization of N_{max} should be improved. Spatial division was conducted using the GOZH model, which divides the study area according to the longitude and latitude. The division process may introduce uncertainty because the geographical environment usually does not have a spatial pattern similar to that of the longitude and latitude. In this study, our primary aim was to propose and verify the idea of borrowing information using ATAK; therefore, only simple and straightforward methods were used for these steps. More advanced and accurate spatial division algorithms should be used in future studies, such as k-means and density-based spatial

clustering of applications with noise (DBSCAN) (Hartigan and Wong 1979, Hahsler *et al.* 2019). Second, the proposed method is a geostatistic interpolation model without combining machine learning and other learning methods. As previous work has proven the potential of machine learning in spatial interpolation (Zhu *et al.* 2020), we will explore how to combine it with GHM to obtain better accuracy.

5. Conclusions

In this study, a Generalized Heterogeneity Model (GHM) was developed to improve the spatial interpolation accuracy of data in large areas. The study space was divided into several strata according to geographical distributions. The spatial dependence within observations in the interpolated stratum was characterized by OK, and the spatial dependence between observations from different strata was characterized by ATAK. The results of the case study demonstrate that GHM had the highest accuracy in terms of MAE and RMSE, compared with other widely used interpolation models, including OK, KED and StK. In addition, the GHM avoided bull's eye patterns and abrupt changes along the strata boundaries.

This paper presents an effective approach for interpolating spatial second-order non-stationary surfaces. We characterized the spatial dependence of different heterogeneous partitions by introducing ATAK, which we hope will inspire future spatial interpolation and prediction. For large-scale regions, both natural and socio-economic variables tend to exhibit spatial second-order non-stationarity, and the GHM method has the potential to effectively interpolate and predict them spatially.

Acknowledgements

We would like to thank the editors and anonymous reviewers for their constructive suggestions and comments for improving this manuscript. We also thank Dr. Maogui Hu for his insightful suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Peng Luo is currently a PhD candidate at the Chair of Cartography and Visual Analytics at the Technical University of Munich, Germany. His research interests include spatial association modeling, social sensing and applied artificial intelligence.

Yongze Song is a Lecturer at Curtin University, Australia and a Fellow of the Royal Geographical Society (with IBG), United Kingdom. His current research interests include geospatial analysis methods, spatial statistics, sustainable development and infrastructure management.

Di Zhu is an Assistant Professor of Geographic Information Science at the University of Minnesota, Twin Cities (UMN). His research aims at generating both theoretical and actionable insights from spatiotemporal data by exploring the frontiers that bridge geospatial analysis, artificial intelligence and social sensing.

Junyi Cheng is currently a PhD candidate at the Institute of Remote Sensing and Geographic Information System, Peking University. His research interests include data mining, the application of big data in public security and information systems.

Liqiu Meng is a professor of Cartography at the Technical University of Munich and a member of German National Academy of Sciences. She is serving as Vice President of the International Cartographic Association. Her research interests include geodata integration, mobile map services, multimodal navigation algorithms, geovisual analytics and ethical concerns in social sensing.

ORCID

Peng Luo http://orcid.org/0000-0002-3680-8509
Yongze Song http://orcid.org/0000-0003-3420-9622
Di Zhu http://orcid.org/0000-0002-3237-6032
Junyi Cheng http://orcid.org/0000-0002-9225-2824
Liqiu Meng http://orcid.org/0000-0001-8787-3418

Data and codes availability statement

The data and codes that support the findings of the present study are available on Figshare at https://doi.org/10.6084/m9.figshare.19604245.

References

Bourennane, H., King, D., and Couturier, A., 2000. Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. *Geoderma*, 97 (3-4), 255–271.

Bowman, M.J., and Esaias, W.E., 1981. Fronts, stratification, and mixing in long island and block island sounds. *Journal of Geophysical Research*, 86 (C5), 4260–4264.

Chen, K., *et al.*, 2020. Land use transitions and urban-rural integrated development: theoretical framework and china's evidence. *Land Use Policy*, 92, 104465.

Chiles, J.P., and Delfiner, P., 2009. *Geostatistics: modeling spatial uncertainty*. Vol. 497. New York, NY: John Wiley & Sons.

Chipman, H.A., George, E.I., and McCulloch, R.E., 1998. Bayesian cart model search. *Journal of the American Statistical Association*, 93 (443), 935–948.

Davies, C.H., *et al.*, 2018. A database of chlorophyll a in Australian waters. *Scientific Data*, 5 (1), 1–8.

De Smith, M.J., Goodchild, M.F., and Longley, P., 2007. *Geospatial analysis: a comprehensive guide to principles, techniques and software tools*. Leicester, UK: Troubador Publishing Ltd.

Elumalai, V., *et al.*, 2017. Spatial interpolation methods and geostatistics for mapping groundwater contamination in a coastal area. *Environmental Science and Pollution Research International*, 24 (12), 11601–11617.

Erickson, R.A., 1983. The evolution of the suburban space economy. *Urban Geography*, 4, 95–121.

Fortin, M.-J., *et al.*, 1996. Quantification of the spatial co-occurrences of ecological boundaries. *Oikos*, 77 (1), 51–60.

Gao, B., *et al.*, 2015. A stratified optimization method for a multivariate marine environmental monitoring network in the Yangtze River estuary and its adjacent sea. *International Journal of Geographical Information Science*, 29 (8), 1332–1349.

Gao, B., *et al.*, 2020. Spatial interpolation of marine environment data using P-MSN. *International Journal of Geographical Information Science*, 34 (3), 577–603.

Geddes, A., *et al.*, 2013. Stochastic model-based methods for handling uncertainty in areal interpolation. *International Journal of Geographical Information Science*, 27 (4), 785–803.
Gong, G., Mattevada, S., and O'Bryant, S.E., 2014. Comparison of the accuracy of kriging and IDW interpolations in estimating groundwater arsenic concentrations in Texas. *Environmental Research*, 130, 59–69.
Goodchild, M.F., 2004. The validity and usefulness of laws in geographic information science and geography. *Annals of the Association of American Geographers*, 94 (2), 300–303.
Goovaerts, P., 1997. *Geostatistics for natural resources evaluation*. New York, NY: Oxford University Press on Demand.
Goovaerts, P., 2010. Combining areal and point data in geostatistical interpolation: applications to soil science and medical geography. *Mathematical Geosciences*, 42 (5), 535–554.
Gotway, C.A., and Young, L.J., 2002. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97 (458), 632–648.
Guan, Q., Kyriakidis, P.C., and Goodchild, M.F., 2011. A parallel computing approach to fast geostatistical areal interpolation. *International Journal of Geographical Information Science*, 25 (8), 1241–1267.
Hahsler, M., Piekenbrock, M., and Doran, D., 2019. dbscan: fast density-based clustering with r. *Journal of Statistical Software*, 91 (1), 1–30.
Hartigan, J.A., and Wong, M.A., 1979. Algorithm as 136: a k-means clustering algorithm. *Applied Statistics*, 28 (1), 100–108.
Hu, M., and Huang, Y., 2020. atakrig: an r package for multivariate area-to-area and area-to-point kriging predictions. *Computers & Geosciences*, 139, 104471.
Hudson, G., and Wackernagel, H., 1994. Mapping temperature using kriging with external drift: theory and an example from Scotland. *International Journal of Climatology*, 14 (1), 77–91.
Hutchings, P., *et al.*, 2022. Understanding rural–urban transitions in the global south through peri-urban turbulence. *Nature Sustainability*, 5 (11), 1–7.
Kyriakidis, P.C., 2004. A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36 (3), 259–289.
Kyriakidis, P.C., and Goodchild, M.F., 2006. On the prediction error variance of three common spatial interpolation schemes. *International Journal of Geographical Information Science*, 20 (8), 823–855.
Lam, N.S.N., 1983. Spatial interpolation methods: a review. *The American Cartographer*, 10 (2), 129–150.
Li, J., and Heap, A.D., 2008. *A review of spatial interpolation methods for environmental scientists*. Canberra, Australia: Geoscience Australia Canberra.
Lichtenstern, A., 2013. Kriging methods in spatial statistics. Munich, Germany.
Likas, A., Vlassis, N., and Verbeek, J.J., 2003. The global k-means clustering algorithm. *Pattern Recognition*, 36 (2), 451–461.
Liu, Y., *et al.*, 2021. Geographical detector-based stratified regression kriging strategy for mapping soil organic carbon with high spatial heterogeneity. *CATENA*, 196, 104953.
Luo, P., *et al.*, 2019. Modeling population density using a new index derived from multi-sensor image data. *Remote Sensing*, 11 (22), 2620.
Luo, P., *et al.*, 2022. Identifying determinants of spatio-temporal disparities in soil moisture of the northern hemisphere using a geographically optimal zones-based heterogeneity model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 185, 111–128.
Luo, P., Song, Y., and Wu, P., 2021. Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level. *GIScience & Remote Sensing*, 58 (5), 756–775.
Ma, T., *et al.*, 2015. Night-time light derived estimation of spatio-temporal characteristics of urbanization dynamics using DMSP/OLS satellite data. *Remote Sensing of Environment*, 158, 453–464.
Matheron, G., 1963. Principles of geostatistics. *Economic Geology*, 58 (8), 1246–1266.

Mitas, L., and Mitasova, H., 1999. Spatial interpolation. *In*: P. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, eds. *Geographical information systems: principles, techniques, management and applications*, vol. 1. New York, NY: Wiley, 481–492.
Oliver, M.A., and Webster, R., 1990. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems* , 4 (3), 313–332.
Preston, R.E., 1966. The zone in transition: a study of urban land use patterns. *Economic Geography*, 42 (3), 236–260.
Sadahiro, Y., 2000. Accuracy of count data transferred through the areal weighting interpolation method. *International Journal of Geographical Information Science*, 14 (1), 25–50.
Song, Y., 2022. The second dimension of spatial association. *International Journal of Applied Earth Observation and Geoinformation*, 111, 102834.
Song, Y., *et al.*, 2018. Segment-based spatial analysis for assessing road infrastructure performance using monitoring observations and remote sensing data. *Remote Sensing*, 10 (11), 1696.
Song, Y., *et al.*, 2020a. An optimal parameters-based geographical detector model enhances geographic characteristics of explanatory variables for spatial heterogeneity analysis: cases with different types of spatial data. *GIScience & Remote Sensing*, 57 (5), 593–610.
Song, Y., *et al.*, 2020b. A spatial heterogeneity-based segmentation model for analyzing road deterioration network data in multi-scale infrastructure systems. *IEEE Transactions on Intelligent Transportation Systems*, 22 (11), 7073–7083.
Song, Y., and Wu, P., 2021. An interactive detector for spatial associations. *International Journal of Geographical Information Science*, 35 (8), 1676–1701.
Tobler, W., 2004. On the first law of geography: a reply. *Annals of the Association of American Geographers*, 94 (2), 304–310.
Wang, J.F., *et al.*, 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *International Journal of Geographical Information Science*, 24 (1), 107–127.
Wang, J.F., Zhang, T.L., and Fu, B.J., 2016. A measure of spatial stratified heterogeneity. *Ecological Indicators*, 67, 250–256.
Wong, T.T., 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48 (9), 2839–2846.
Yoo, E.H., and Kyriakidis, P.C., 2006. Area-to-point kriging with inequality-type data. *Journal of Geographical Systems*, 8 (4), 357–390.
Zhang, Z., Song, Y., and Wu, P., 2022. Robust geographical detector. *International Journal of Applied Earth Observation and Geoinformation*, 109, 102782.
Zhu, D., *et al.*, 2020. Spatial interpolation using conditional generative adversarial neural networks. *International Journal of Geographical Information Science*, 34 (4), 735–758.

Appendix A

$$\begin{aligned} E\left(\hat{Z}_0 - Z_0\right) &= E\left(\sum_{i=1}^{n}\lambda_i Z_i + \sum_{j=n+1}^{n+k}\lambda_j Z_j - Z_0\right) \\ &= \sum_{i=1}^{n}\lambda_i E(Z_i) + \sum_{j=n+1}^{n+k}\lambda_j E(Z_j) - E(Z_0) \\ &= \left(\sum_{i=1}^{n}\lambda_i - 1\right) * m_{s1} + \left(\sum_{j=n+1}^{n+k}\lambda_j * m_{s2}\right) \end{aligned} \tag{A1}$$

where m_{s1}, m_{s2} are the expectations of the variable inside and outside the interpolation stratum, respectively. In this study, m_{s1} is the mean value of observations in the interpolated stratum, and m_{s2} is the mean value of observations outside the interpolated stratum.

Appendix B

The estimation error is transformed into the following equation using the residues:

$$\hat{Z}_0 - Z_0 = \sum_{i=1}^{n} \lambda_i (Z_i - m_{s1}) + \sum_{j=n+1}^{n+k} \lambda_j (Z_j - m_{s2}) - (Z_0 - m_{s1}) = \sum_{i=1}^{n} \lambda_i R_i + \sum_{j=n+1}^{n+k} \lambda_j R_j - R_0 \tag{B1}$$

where R_i, R_j, and R_0 are the residues of Z_i, Z_j, and Z_0 after removing the expectations, respectively.

$$\begin{aligned}
\delta_E^2 &= \mathrm{var}(\hat{Z}_0 - Z_0) = \mathrm{var}\left[\left(\sum_{i=1}^{n} \lambda_i R_i + \sum_{j=n+1}^{n+k} \lambda_j R_j\right) - R_0\right] \\
&= \mathrm{var}\left(\sum_{i=1}^{n} \lambda_i R_i + \sum_{j=n+1}^{n+k} \lambda_j R_j\right) - 2\mathrm{Cov}\left[\left(\sum_{i=1}^{n} \lambda_i R_i + \sum_{j=n+1}^{n+k} \lambda_j R_j\right), R_0\right] \\
&+ \mathrm{var}(R_0) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{Cov}_{inside}(R_i, R_j) + \sum_{i=n+1}^{n+k}\sum_{j=n+1}^{n+k} \lambda_i \lambda_j \mathrm{Cov}_{outside}(R_i, R_j) \\
&+ 2\sum_{i=1}^{n}\sum_{j=n+1}^{n+k} \lambda_i \lambda_j \mathrm{Cov}_{between}(R_i, R_j) - 2\sum_{i=1}^{n} \lambda_i \mathrm{Cov}_{inside}(R_i, R_0) \\
&- 2\sum_{j=n+1}^{n+k} \lambda_j \mathrm{Cov}_{between}(R_j, R_0) + \mathrm{var}(R_0)
\end{aligned} \tag{B2}$$

where Cov_{inside} is the spatial covariance function of the stratum to be interpolated, $Cov_{outside}$ is the spatial covaiance function of all strata outside the stratum to be interpolated, and Cov_{inside} is the spatial covariance function between the interpolated stratum and other strata.

To minimize the δ_E^2 as well as achieve the unbiased estimation (Equations (3) and (A1)), the Lagrange multiplied was introduced to solve this optimization problem. The built Lagrange function is as follows:

$$J = \delta_E^2 + 2\ \mathrm{L}\left[\left(\sum_{i=1}^{n} \lambda_i - 1\right) * m_{s1} + \left(\sum_{j=n+1}^{n+k} \lambda_j * m_{s2}\right)\right] \tag{B3}$$

There are three unknown variable sets: λ_i ($i = 1, 2, \ldots n$), λ_i ($i = n+1, n+2, \ldots n+k$), and L. The matrix includes totally $n + k + 1$ unknown values. Solving the matrix using the Lagrange multiplier as follows:

$$\begin{cases}
\dfrac{\partial J}{2\partial \lambda_i} = \displaystyle\sum_{j=1}^{n} \lambda_j \mathrm{Cov}(R_i, R_j) + \sum_{j=n+1}^{n+k} \lambda_j \mathrm{Cov}(R_i, R_j) + L = \mathrm{Cov}(R_0, R_i)(i = 1, 2 \ldots n) \\
\dfrac{\partial J}{2\partial \lambda_j} = \displaystyle\sum_{j=n+1}^{n+k} \lambda_j \mathrm{Cov}(R_i, R_j) + \sum_{j=1}^{n} \lambda_j \mathrm{Cov}(R_i, R_j) + L = \mathrm{Cov}(R_0, R_i)(i = n+1, n+2 \ldots n+k) \\
\left(\displaystyle\sum_{i=1}^{n} \lambda_i - 1\right) * m_{s1} + \left(\sum_{j=n+1}^{n+k} \lambda_j * m_{s2}\right) = 0
\end{cases} \tag{B4}$$

These three equations are transformed as the matrix to be clearly understood, as follows:

$$\begin{bmatrix}
R_{1,1} & \ldots & R_{1,n+k} & m_{s1} \\
R_{2,1} & \ldots & R_{2,n+k} & m_{s1} \\
\ldots & & & \ldots \\
R_{n+k,1} & \ldots & R_{n+k,n+k} & m_{s2} \\
m_{s1} & \ldots & m_{s2} & 0
\end{bmatrix}
\begin{bmatrix}
\lambda_1 \\ \lambda_2 \\ \ldots \\ \lambda_{n+k} \\ L
\end{bmatrix}
=
\begin{bmatrix}
R_{1,0} \\ R_{2,0} \\ \ldots \\ R_{n+k,0} \\ m_{s1}
\end{bmatrix} \tag{B5}$$

where $R_{i,j}$ represents the $Cov(R_i, R_j)$.