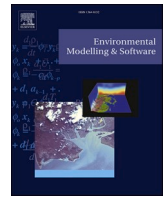




Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: <http://www.elsevier.com/locate/envsoft>

The spatial statistic trinity: A generic framework for spatial sampling and inference

Jinfeng Wang^{a,*}, Bingbo Gao^b, Alfred Stein^c^a State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences, A11 Datun Road, Beijing, 100101, China^b College of Land Science and Technology, China Agriculture University, Tsinghua East Road, Beijing, 100083, China^c Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7514, AE Enschede, the Netherlands

ARTICLE INFO

Keywords:

Population and sample
 Spatial autocorrelation (SAC)
 Spatial stratified heterogeneity (SSH)
 Variable and random variable
 Spatial statistic trinity (SST)

ABSTRACT

Geospatial referenced environmental data are extensively used in environmental assessment, prediction, and management. Data are commonly obtained by nonrandom surveys or monitoring networks, whereas spatial sampling and inference affect the accuracy of subsequent applications. Design-based and model-based procedures (DB and MB for short) both allow one to address the gap between statistical inference and spatial data. Creating independence by sampling implies that DB may neglect spatial autocorrelation (SAC) if the sampling interval is beyond the SAC range. In MB, however, a particular sampling design can be irrelevant for inferential results. Empirical studies further showed that MSE (mean squared error) values for both DB and MB are affected by SAC and spatial stratified heterogeneity (SSH). We propose a novel framework for integrating SAC and SSH into DB and MB. We do so by distinguishing the spatial population from the spatial sample. We show that spatial independence in a spatial population results in independence in a spatial sample, whereas SAC in a spatial population is reflected in a spatial sample if sampling distances are within the range of dependence; otherwise, SAC is absent in the spatial sample. Similarly, SSH in a population may or may not be inherited in data, and this depends on the sampling method. Thus, the population, sample, and inference constitute a so-called spatial statistic trinity (SST), providing a new framework for spatial statistics, including sampling and inference. This paper shows that it greatly simplifies the choice of method in spatial sampling and inferences. Two empirical examples and various citations illustrate the theory.

1. Introduction

Spatial sampling and inference are widely used in the survey and assessment of soil, atmosphere, and water environments (Gao et al., 2015; Mindham and Tych, 2019; Wang and Xuan, 2020; Wang et al., 2013b; Chen et al., 2019). Inappropriate application of the methods, however, may introduce bias in modelling and analysis. The model-based (MB) and the design-based (DB) statistical procedures are two common ways to address this problem. DB originated from Bowley (1906), and it was further shaped in the 1930s (e.g., Neyman, 1934), whereas MB originated from Kiaer (1896), and it was shaped by Fisher, 1922; in Section 2) and later by Krige (1951) and Matheron (1963, 1971). The two statistical frameworks constitute two mainstream approaches in spatial statistics (Fig. 1; Cassel et al., 1977; de Gruijter and ter Braak, 1990; Brus and Guijter, 1997). In DB, the population is

regarded as fixed, and stochasticity in the data is introduced by random sampling; estimation procedures are based on the known probabilities of sampling. Spatial variation within a population, mainly described as either spatial autocorrelation (SAC), or spatial stratified heterogeneity (SSH), is of little concern in DB (de Gruijter and ter Braak, 1990). Conversely, in MB, the population is regarded as a single realization of a joint distribution model, and predictions are based on the model chosen to represent the population. Joint design- and model-based frameworks (DMFs) provide a general basis for statistical sampling and inference.

The distinction between DB and MB raises some concerns: (1) Will the spatial variation (SAC and SSH) of a population affect the MSE of estimation when applying DB? (2) Can information contained in the sampling design be used to improve the MSE of estimation when applying MB? (3) There is some overlap between the two approaches when aiming at the same target (de Gruijter et al., 2006, p.68). To

* Corresponding author.

E-mail addresses: wangjf@lreis.ac.cn (J. Wang), gaobingbo@cau.edu.cn (B. Gao), a.stein@utwente.nl (A. Stein).

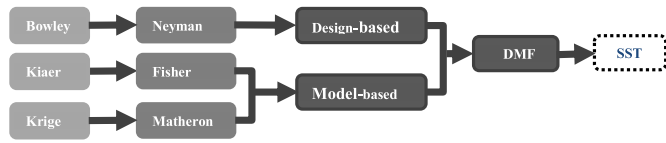


Fig. 1. Evolution of the design- and model-based frameworks (DMFs).

Table 1
Types of sampling strategies updated from Brus and Guijter (1997).

		Values at given location	
		Fixed	Random
Sample locations	Fixed	Fully deterministic	Model-based
	Random	Design-based	Hybrid

address the first issue, we note that in DB, stratified sampling is favored over simple random sampling if the population exhibits SSH (Cochran, 1978; Wang et al., 2016). For the second issue, it is well known that in MB, the empirical error of kriging interpolation, i.e., the difference between the predicted value and the true value at a prediction point, varies with the sampling design (Minasny and McBratney, 2016). For the third issue, we note that overlapping makes it difficult for users to make the best choice of approach. Model-assisted sampling and hybrid sampling overcome part of the limitations of the DMF (Särndal et al., 1992; Sterba, 2009). In hybrid sampling, the spatial variation of a population can be included in the joint stochastic distribution model, and a sampling design of any random sample can be used in the estimation of parameters. Accordingly, Brus and Guijter's (1997) table of sampling strategies can be updated as shown in Table 1.

The objective of this paper is to explore the root of the inconsistency between theoretical assumptions and empirical studies for both DB and MB. That is, DB neglects the spatial variation of a population and MB neglects the sampling design, while in empirical studies, both of the neglects influence the errors of their estimates from the real value. Consequently, we propose a unified framework named the spatial statistic trinity (SST) that will greatly simplify the model and method choice in the spatial sampling and spatial inference of environments.

To address the problems above, the remainder of this paper is organized as follows. Section 2 refines the common standard of performance for various samplings and inferences; Section 3 proposes the generic SST framework to address the problems (arisen in Section 1)

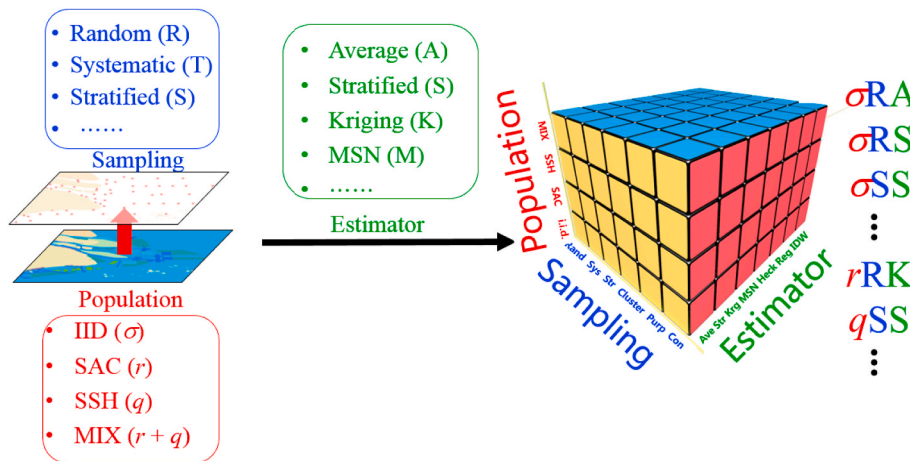


Fig. 2. Spatial statistic trinity. i.i.d., SAC, SSH and MIX in red color are four types of population; Random, Systematic and Stratified in blue color are sampling methods, representing simple random sampling, systematic sampling, stratified random sampling respectively; Average, Stratified, Kriging, MSN in green color are estimators, representing estimator for simple random sampling, estimator for stratified random sampling, Kriging and MSN; each square in the magic cube represents a combination of a population, sampling method and estimator; the letters in the rightmost column are examples of the combination, with the first letter representing the population, the middle letter representing the sampling method and the third letter representing the estimator. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

using the standard of performance (set in Section 2). Then, SST is compared to the existing DMF and others in Section 4. Finally, we illustrate the key concepts of SST using empirical studies in Section 5, and we draw conclusions in Section 6.

These Sections and Sub-sections are supported and logically linked by sequential figures as well: study background (Fig. 1) → motivation of SST with an example (Figs. 2 and 3) → connections between the components in SST using a data flow (Fig. 4) → computation issues (Fig. 5) → model efficiency comparison (Fig. 6) → SST model choice for practical applications (Fig. 7). Key concepts in SST are empirically illustrated (Figs. 8–10).

2. The performance of spatial sampling and statistical inference

The gold standard of the performance of an estimator or predictor Ψ is its difference from the true value Ψ_0 :

$$|\Psi - \Psi_0| \tag{1}$$

where Ψ is a parameter of a population \mathfrak{R} or superpopulation \mathbb{R} . A superpopulation refers to a collection of random variables (one for each location), and a population is one of the realizations of the superpopulation. In practice, however, Ψ_0 is unknown, and usually only a single realization of a superpopulation or one sample \mathfrak{S} of a population is available. The performance of an estimator $\hat{\Psi}$ is then measured via the theoretical expectation of the p -MSE under the assumption of many instances of sampling of the target population by a sampling scheme p (Equation (2a)). Likewise, the performance of a predictor $\tilde{\Psi}$ is measured by the theoretical expectation of the ξ -MSE under the assumption of many instances of realizations of the superpopulation ξ (Equation (2b)) or the joint p - and ξ -MSE (Equation (2c)):

$$\text{MSE}_p \Psi(\mathfrak{S}) = E_p(\Psi - \Psi_0)^2 \tag{2a}$$

$$\text{MSE}_\xi \Psi(\mathfrak{R}) = E_\xi(\Psi - \Psi_0)^2 \tag{2b}$$

$$\text{MSE}_{\xi p} \Psi = E_{\xi p}(\Psi - \Psi_0)^2 \tag{2c}$$

where the MSE refers to the mean square error and Ψ may be either an estimator $\hat{\Psi}$ or a predictor $\tilde{\Psi}$. The MSE reduces to a variance in the case of an unbiased estimation:

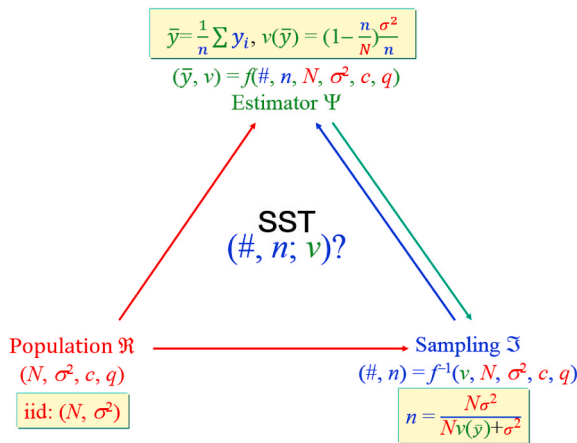


Fig. 3. An example of an SST triple with an i.i.d. population (in the box). # represents the sampling method, e.g., random, systematic, kriging; n is the number of sample units; v is the error of estimation; N is the number of units in the population; σ^2 is the dispersion variance; c is the spatial autocorrelation; and q is the spatial stratified heterogeneity.

$$E_p \Psi(\mathfrak{S} | \mathfrak{R}) = \sum_{\mathfrak{S}=1}^{\mathfrak{R}} p(\mathfrak{S}) \Psi(\mathfrak{S} | \mathfrak{R}) \stackrel{\text{def}}{=} M(\mathfrak{R}) \quad (3a)$$

$$E_{\xi} \Psi(\mathfrak{R} | \mathbb{R}) = \sum_{\mathfrak{R}=1}^{\mathbb{R}} p(\mathfrak{R}) \Psi(\mathfrak{R} | \mathbb{R}) \stackrel{\text{def}}{=} M(\mathbb{R}) \quad (3b)$$

In this study, we focus on an unbiased estimation. For such a population parameter as the spatial mean, p -inference fails to account for the SAC in a population (Särndal, 1978; de Gruijter and ter Braak, 1990). For a superpopulation parameter, such as the model mean, ξ -inference only concerns the distances between sampling sites; in contrast, it fails to account for their specific locations. Note that the theoretical measures in statistics, originating from gambling based on random and repeated sampling (RRS) and widely applied in experimental sciences, seem too far from the reality in geoscience, where both random and repeated sampling are rare.

3. The spatial statistic trinity (SST) framework

3.1. Motivation of SST

In MB, sampling is done on a population variable, whereas spatial data refer to a sample (Cochran, 1977; Anselin, 1988; Cressie, 1992; Cressie, 1990; Haining, 2003). While in DB one assumes a fixed population, both the population and sample can be referred to as spatial data. To improve estimation for spatial data, we have to deal with SAC, defined as the coincidence of locational and attribute similarities (Tobler, 1970; Matheron, 1963; Anselin, 1988; Goodchild, 2004; Sui, 2004), although SAC is not explicitly included in DB inference. SAC is a

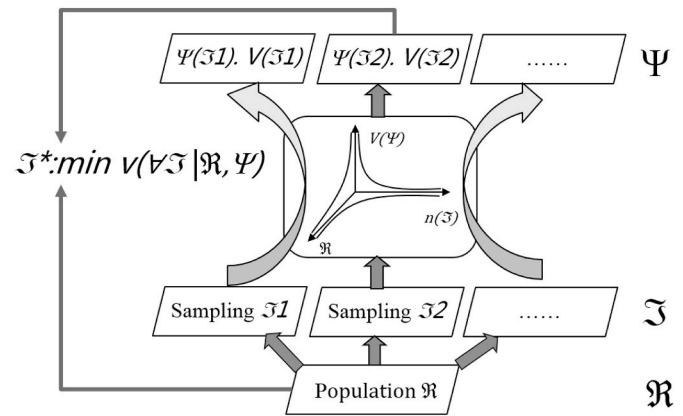


Fig. 5. Optimal sampling for a population. (v stands for variance).

property of either the population or the sample, with the former being independent of the latter. Kriging interpolation will be useful if the population has SAC, particularly if distances between sampled sites are less than the range of spatial dependence.

Instead of the (p, t) spatial statistical dual (Brus and Guijter, 1997; Brus, 2019), where p represents a sampling design and t represents an estimator, we put forward an SST $(\mathfrak{R}, \mathfrak{S}, \Psi)$ composed of a population \mathfrak{R} , sampling design \mathfrak{S} , and estimator Ψ . Each of the three dimensions has many options; consequently, there are more than 100 combinations of triples (Fig. 2). Only a few of these are efficient, however, where the population properties, assumption of estimators, and sampling conditions match. Fig. 3 gives a simple example of SST: an i.i.d. population, random sampling, and sample mean in which the triple of the population parameters (number of units N and variance σ^2), of the statistic parameters (sample mean \bar{y} and its variance v), and of the sampling parameters (sample size n) are connected by a function having an analytic solution.

3.2. SST components and their relationships

Fig. 4 illustrates the spatial data flow in SST from a generator, usually a spatial process or superpopulation \mathbb{R} , to its single realization over a geographical space (population \mathfrak{R}), which is observed by a single sampling (\mathfrak{S}). We are interested in predicting the values at unsampled sites, the spatial mean of the population \mathfrak{R} , or even the parameters of the superpopulation \mathbb{R} based on the observed sample \mathfrak{S} and an estimator or predictor Ψ under specific assumptions. Drawing a sample from a population, the values of the sample are the result of data propagation from superpopulation \mathbb{R} to population \mathfrak{R} under sampling scheme \mathfrak{S} . The spatial variation of a superpopulation, such as i.i.d., SAC, or SSH, is inherited by the population (Haining, 1988, p.575; Eqs. (1.1), (1.2); Cressie, 1993, pp.13–15). These properties, however, may or may not be inherited by the sample due to the sampling method (Isaaks and

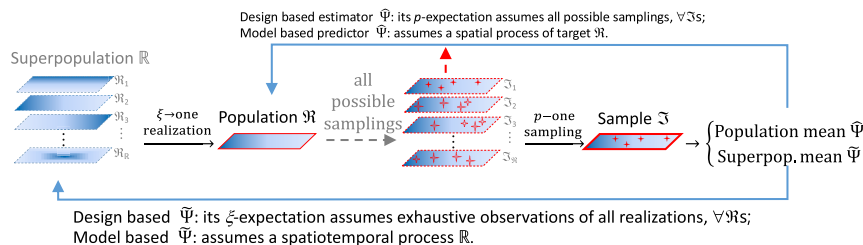


Fig. 4. Spatial data flow. Solid arrows denote observed information flow; dashes and grey words denote imagined realizations and samplings. The stochastic properties of an observed population generated by a superpopulation and a sample design from a population to a sample are denoted by ξ and p , respectively. The population mean $\hat{\Psi}$ is called the *spatial mean* and the superpopulation mean $\hat{\Psi}$ the *model mean*; + stands for sample sites.

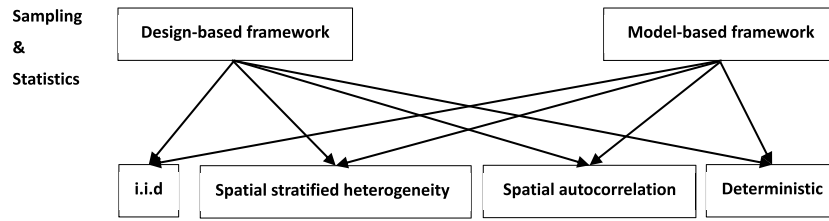


Fig. 6. Efficiency of spatial sampling and statistics for different populations. (the shorter the length of the arrow, the higher the efficiency of the framework).

Population \mathfrak{R}		Stratified Heterogeneity (Geodetector q)	
		Absent ($q = 0$)	Present ($q \neq 0$)
Spatial Autocorrelation (Moran I; Semivariogram)	Absent ($I = 0$)	Simple Ave	
	Present ($I \neq 0$)	Kriging	

Fig. 7. Model selection according to SAC (shaded cells) and SSH (cells partitioned by curves) for spatial mean (rows indicate whether the population exhibits SAC, i.e., $I = 0$ or $I \neq 0$, respectively); columns indicate whether the population presents SSH, i.e., Geodetector $q = 0$ or $q \neq 0$, respectively; and the $q \neq 0$ column is further divided into two columns according to whether the sample covers all strata; dots stand for sample units; Sample Ave is the estimation method of simple random sampling; Sandwich is proposed by Wang et al., [2013a]; MSN is by proposed by Wang et al., [2009]; Heckman’s correction methods are proposed by Heckman [1979]; Bshade is proposed by Wang et al., [2011]; SPA is proposed by Wang et al., [2012].

Srivastava, 1989, chapter 7; Haining, 1988, p.576), and they may or may not be utilized by the chosen estimator. Because of the propagation and inheritance of the data properties (Table 2), we coin either spatial i.i.d. or SAC or SSH for a population instead of for a superpopulation, as linear regression modelling is interpreted under the framework of population and sample without recourse to the word of superpopulation (Gujarati and Porter, 2009, p.35, 43). When the distinction between population and superpopulation causes little difference, for simplification, we use population to refer to both population and the mechanism generating the population, i.e., superpopulation. Consequently, an estimation using the sample data \mathfrak{S} and an estimator Ψ are determined by the sampling trinity SST ($\mathfrak{R}, \mathfrak{S}, \Psi$). SST considers the uncertainties in superpopulation modelling and population sampling, as well as the match between the assumption of a model or estimator for inference and the property of a population.

The elements and linkages of the spatial data flow in Fig. 4 and Table 2 are further explained in the following subsections.

3.2.1. Population (\mathfrak{R}) and estimator (Ψ)

A spatial process refers to a superpopulation, where its properties are inherited by its populations. For example, disease risk (superpopulation \mathbb{R}) is to be distinguished from an exhaustive observation (population \mathfrak{R}) of disease incidence or prevalence if we are concerned with its causation. One commonly assumes that a disease occurrence follows a Poisson process (Haining, 2003, p.308), $O_i \sim \text{Poisson}(E_i r_i)$, where O_i is the observed number of deaths (population) in the i -th subarea, and E_i denotes the expected number of deaths from the disease. Then, the relative risk of dying from the disease r_i , with $r_i = O_i/E_i$ as its maximum likelihood estimator, is a superpopulation parameter that can

be estimated and mapped for each subarea. The risk r_i may be further modeled by a prior distribution or covariates (Banerjee et al., 2015). Based on different perspectives, population and superpopulation may be defined as interchangeable. If it concerns individual locations, then there are two concepts for estimation, which are as follows: $p(\text{disease occurrence} | \text{locations})$ and $p(\text{locations} | \text{disease occurrence})$. These concepts represent the probability of the diseases $p(y)$ and the probability function of the locations $p(\text{locations})$. They are related to each other through Bayes’ law, that is, $p(y | x) = p(x | y)p(y)/p(x)$. Commonly, the term $p(x)$ is ignored because it is irrelevant for y , and we have that $p(y | x) \sim p(x | y)p(y)$.

The spatial mean of a population \mathfrak{R} and the model mean of a superpopulation \mathbb{R} are defined as

$$M(\mathfrak{R}) = \frac{1}{N} \sum_{i=1}^N z(i|\mathfrak{R}), \quad (4a)$$

$$M(\mathbb{R}) = \frac{1}{\mathbb{R}} \sum_{\mathfrak{R}=1}^{\mathbb{R}} M(\mathfrak{R}) \quad (4b)$$

respectively, where $\mathfrak{R} = 1, \dots, \mathbb{R}$, and \mathbb{R} stands for either a superpopulation or the number of all possible populations of a superpopulation. Both $M(\mathfrak{R})$ and $M(\mathbb{R})$ can be estimated by a weighted sample mean:

$$m(\mathfrak{R}) = \sum_{i=1}^n w(i)z(i|\mathfrak{R}) \quad (5)$$

where $w(i)$ is adopted by a sampling design (Horvitz and Thompson, 1952; Särndal et al., 1992), such as $1/n$ if it is simple random sampling, or it is calibrated if the spatial population (variation) is modeled.

3.2.2. Population (\mathfrak{R}) and sampling (\mathfrak{S})

The spatial variation of a population may or may not be reflected in a sample. A population \mathfrak{R} or superpopulation \mathbb{R} can be i.i.d., such as a Poisson process or white noise; deterministic, such as a plane or sinuous time series; or hybrid—that is, between completely random and completely deterministic—such as an SSH or SAC. Given a specific estimator Ψ and sampling method \mathfrak{S} , the MSE varies with the properties of the target populations (Table 3). We consider three cases below.

Case 1. Deterministic population (DP). If a population is fully deterministic, such as the path of a bouncing ball, it follows Newton’s law Ψ . From a low number of sampling points of the path, one can reconstruct the exact path (population) using Newton’s law as an estimator (Isaaks and Srivastava, 1989, p.199). A second example is the detail of urban planning. The third example is the population y , which is determined by the covariate X , such as the first soil map for the United States, which is strongly influenced by geology and rivers (Whitney, 1909); soil-landscape models allow for the prediction of soil properties based on landscape position (Branham, 1989); pollution emissions by vehicles differ between steady speed and unsteady, allowing for the prediction of air pollution along a road. The fourth example is the temporal transmission of a communicable disease, which can be well reflected by the temporal process of the susceptible-exposed-infective-recovered (SEIR)

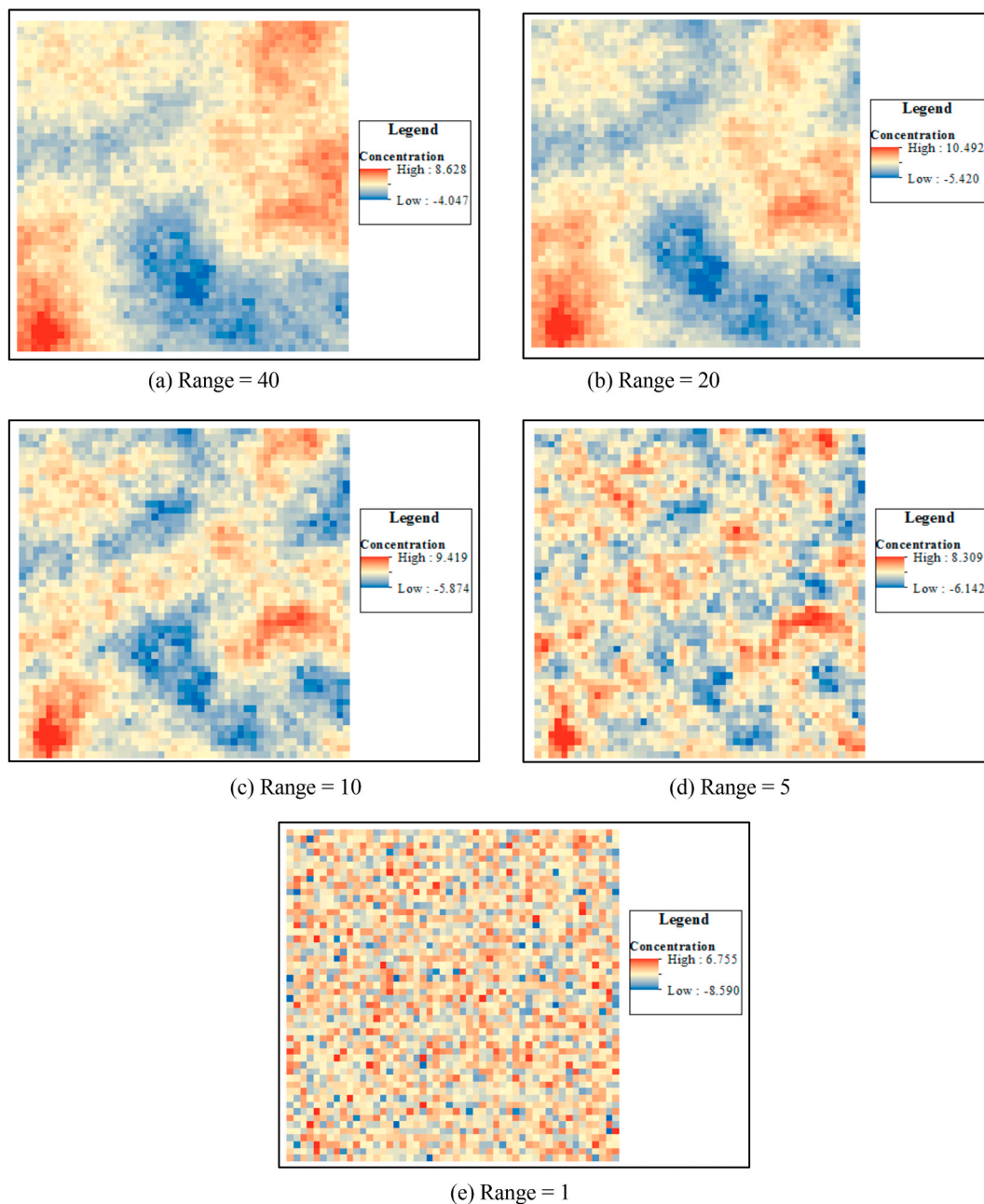


Fig. 8. Five simulated populations with different SAC values (the target value is the Pb concentration in soil with ppm as the unit).

model. A few sample units randomly drawn from the SEIR process can calibrate the SEIR model to simulate the exact time series of an infectious disease (Wang et al., 2006). SAC can be derived from the spatial version of SEIR and then used for spatial interpolation (Angulo et al., 2013; Kolovos et al., 2013).

Case 2. Semi-random population (SRP). For SSH populations, such as climate zones, the annual mean temperature of an area can be estimated using a stratified sampling (Cochran, 1977). If a population exhibits SAC, with a known shape of the variogram—which can be calibrated by a sample even if it is i.i.d. and small—then kriging can be implemented (Matheron, 1963; Rodriguez-Iturbe and Mejia, 1974; Haining, 2003). Context effects and uncertainty can then be observed (Kwan and Schwanen, 2018).

Case 3. An i.i.d. population (IID). Against a population containing SAC or SSH, an i.i.d. population looks like a white noise picture, which may be generated by a spatially independently and identically

distributed superpopulation. An example is a piece of farmland of seedlings that was sowed with problematic corn seeds (only 80% can sprout due to variety or spoiling, for example). It is virtually impossible to reconstruct the i.i.d. population using a finite sample and even advanced estimators. There is no better approach than random sampling for surveying an i.i.d. population; spatial interpolation is impossible in this case (Table 3).

A population \mathfrak{R} or superpopulation \mathfrak{R} could be estimated by a mechanistic model with little data as in case 1, a weak model with much data as in case 3, or a balance between the model and data as in case 2. Any choice of estimation method depends on the properties of the population, as well as the available sample. The properties of a population or superpopulation may be known in advance via the general and specific knowledge of the process (Christakos, 2005), study area, relevant determinants, or prior exploratory sampling. If no prior knowledge of the target population is available, simple random or systematic

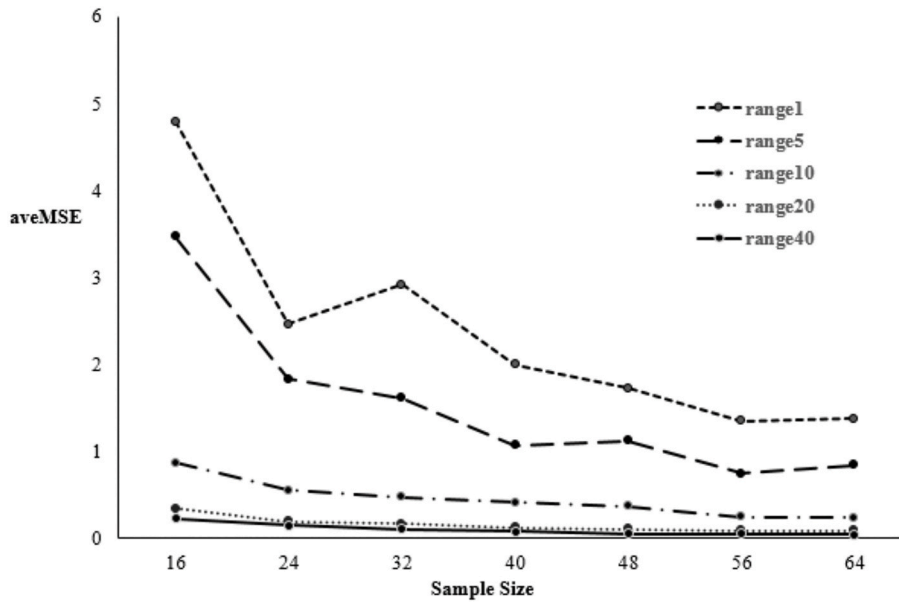


Fig. 9. Estimation errors of simulated datasets (Range1 is the result for population with Range = 1, Range5 for Range = 5, Range10 for Range = 10, Range20 for Range = 20, and Range40 for Range = 40).

sampling is applicable, although this is done at the possible cost of inefficiency.

3.2.3. Sampling (\mathfrak{S}) and estimator (Ψ)

Sampling is usually followed by applying a method to make an inference of a population or superpopulation. By the analytical solution of a sampling design, we mean that the variance of an estimator is a function of sample size, and vice versa. For example, the variances of the superpopulation mean and population mean of a simple random sampling are (Ripley, 1981; Haining, 1988; Griffith et al., 1994)

$$v(\mathbb{R}) = (1 + r) \left(1 - \frac{n}{N}\right) \frac{S^2}{n} \tag{6a}$$

$$v(\mathbb{R}) = (1 - r) \left(1 - \frac{n}{N}\right) \frac{S^2}{n} \tag{6b}$$

respectively, where n is the sample size, $\frac{n}{N}$ is the sampling fraction, S^2 is the variance of the population, and r is the average of pairwise correlations of the study area computed by the covariance between two randomly selected points. For an i.i.d. population, $r = 0$ in (6a, 6b), and the right side becomes the same as the DB one. After a simple transform, the sample size n is a function of the required variance of the sample mean. An effective sample size is $n' = n/(1 + r)$ for an SAC superpopulation and $n' = n/(1 - r)$ for an SAC population (Griffith, 2005).

For most estimators (e.g., kriging predictors), there are no analytic solutions to find the optimal sampling design given their variances; thus, they must be found by simulations. As Fig. 5 shows, for a given population \mathbb{R} and sample size, how simple random sampling is carried out. Sample $\mathfrak{S}1$ is used by estimator Ψ to estimate the spatial mean $\hat{m}(\mathfrak{S}1)$ with variance $v(\mathfrak{S}1)$. Another sampling $\mathfrak{S}2$ is carried out to obtain $\hat{m}(\mathfrak{S}2)$ and $v(\mathfrak{S}2)$. This process is repeated until all combinations C_N^n are enumerated, where N and n are the numbers of units in the population and sample, respectively. The optimal sampling is the sampling that results into the smallest variance. The population mean \hat{m} can be estimated by either a DB estimator ($\Psi(v\mathfrak{S})|\mathbb{R}$) of an average over all samplings \mathfrak{S} s in one realized population \mathbb{R} or an MB predictor ($\Psi(v\mathbb{R})|\mathfrak{S}$) of an average over many realized populations \mathbb{R} s giving one sampling \mathfrak{S} .

3.3. Model selection with SST

Fig. 4 shows that the properties of a spatial sample are the result of information propagation from a process, that is, a superpopulation, to a population; the population is a realization of the superpopulation and is then captured by the sample and operated in an estimator or predictor. SAC and SSH are generic features of a spatial process or population. They may or may not be reflected by a sample (Table 4). If the superpopulation is i.i.d., the population will be i.i.d. In most cases, whereas autocorrelation in a superpopulation will be preserved in its populations and be present in the observed spatial data if the distance between sampling units is within the SAC range. The parameters of a population or superpopulation may be obtained from priors like Newton's law, previous surveys or a survey of a similar attribute in the same area, or surveys of the attribute in a similar area. In practice, SAC can be tested by Moran's I (Moran, 1950) or semivariogram (Matheron, 1963); SSH can be tested by q statistic (Wang et al., 2016); and a population is i.i.d. if it is neither SAC nor SSH.

Many sampling methods and many estimators can be used. Hence, any sampling and estimation can be one of the combinations of the triples (Fig. 2). Table 5 and Fig. 6 illustrate the efficiency of the SST. Estimation method selections according to SAC and SSH are illustrated in Fig. 7. For an i.i.d. population, no approach is better than simple random sampling for inferring about a parameter. When a population only presents SAC without SSH, kriging is preferred. For an SSH population, stratified sampling or Sandwich (Wang et al., 2013a) is superior to simple random sampling because it requires fewer sample units to reach a given inference precision. When both SAC and SSH exist in a population, the population is first partitioned according to its SSH, and it can then be estimated by kriging in strata (Stein et al., 1988; Goovaerts, 1997) or more precisely by the mean of the surface with stratified non-homogeneity (MSN) (Wang et al., 2009; Gao et al., 2020). If the sample cannot cover all strata with an SSH population, bias correction should be adopted, such as the correction method of Heckman (1979) for an SSH population and Bshade (Wang et al., 2011) and SPA (Wang et al., 2012) for a population that presents both SAC and SSH.

In summary, DB (p, t) and MB (ξ) (Brus and Guijter, 1997, p.5) can be integrated by the SST ($\mathbb{R}, \mathfrak{S}, \Psi$). Compared with a complicated DMF-based decision tree for choosing between DB and MB sampling and statistical strategies (Brus and Guijter, 1997, Figure 11), the SST covers

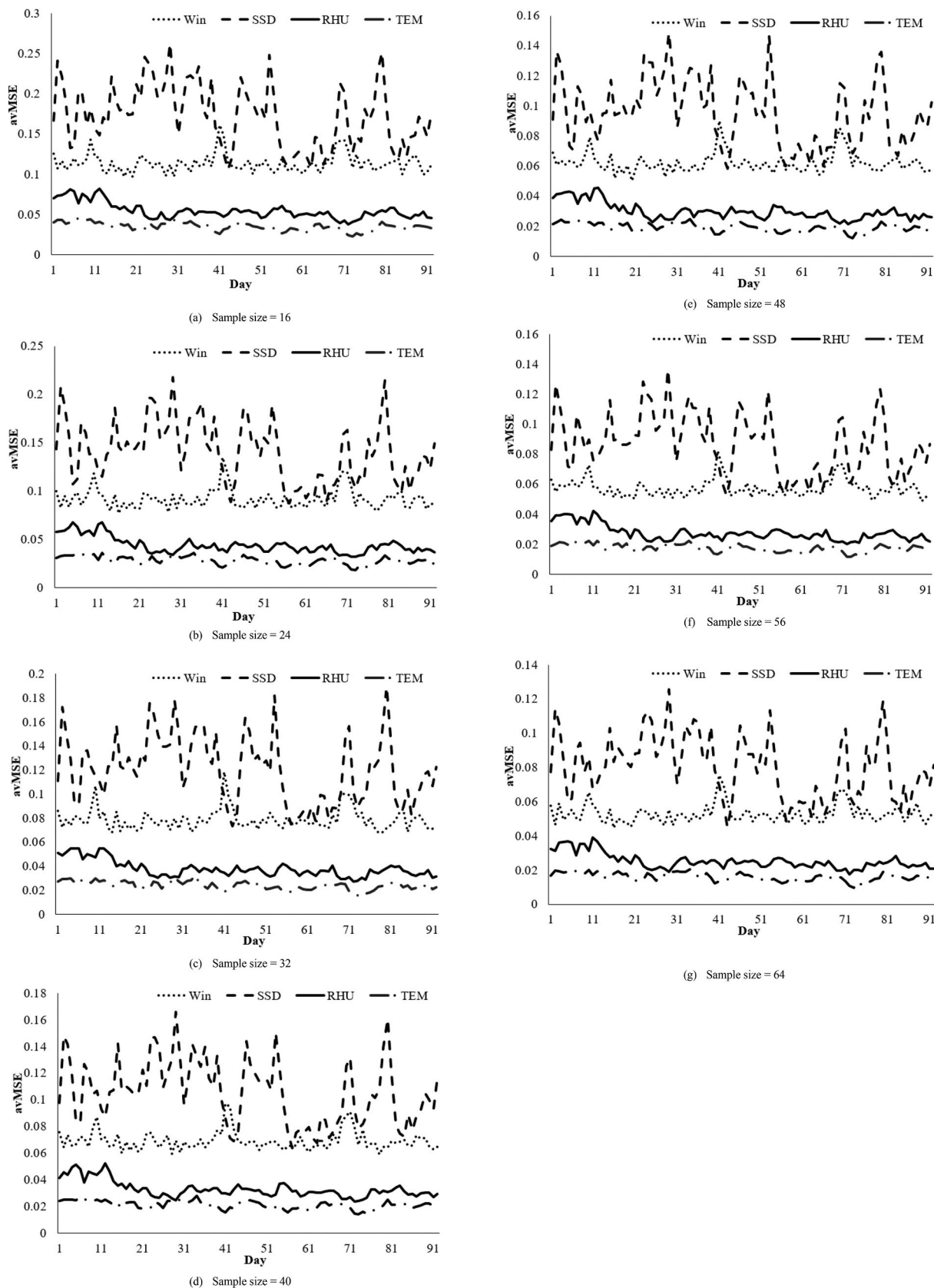


Fig. 10. Estimation error of a climate dataset.

Table 2
Propagation of properties in spatial data.

SAC; SSH; i.i.d.		Then		
		Superpopulation	Population	Sample
If	Superpopulation	–	Y	Y/N
	Population	Y	–	Y/N
	Sample	Y/N	Y/N	–

the essentials of spatial statistics from sampling choice to statistical inference (Table 5).

4. Reinterpretation of key concepts under the SST

In SST, the MSE is determined by the sampling design, estimator, and population properties. This is different from DMF, which only admits two of the triples. Following an influential paper interpreting DMF (Brus and Guijter, 1997), we reinterpreted some key concepts under SST. Table 6 lists the key notations and definitions, which are compared with

Table 3
Conditions for spatial interpolation based on spatial autocorrelation.

Interpolation based on SAC		Sampling	
		i.i.d.	SAC
Population	i.i.d. SAC	N/A Implementable	N/A Implementable

those in authoritative statistics books.

We have outlined the properties below.

R1. Determinants of the estimates. In SST, prediction, estimation and their variances are determined by SST (see Fig. 2). This contrasts with DMF, which considers MB to be unaffected by the sampling design and DB to be unaffected by spatial variation. However, as an MB method, kriging has an empirical error that varies according to the sampling scheme. Random and stratified sampling have developed as DB methods, and the variances of the sample means are proportional to the variance and SSH of the population.

R2. ξ -Unbiasedness and minimum ξ -variance. In SST (Fig. 2), the MSE is the gold standard for assessing the efficiency of spatial sampling and of an estimator. This is determined by SST—that is, the population property, sampling plan, and estimator—regardless of whether a DB or MB method is used. In DMF, ξ -unbiasedness and minimal ξ -variance are not that useful, because in MB, there are no restrictions on the selection of the sampling locations.

R3. Random variable. In SST, the location selection function $I(x)$ is a random variable, which equals 1 if site x is selected and 0 otherwise; x_i refers to a location i , which is fixed once selected. Thus, it is considered a variable but not a random variable. This contrasts with DMF, where the i -th location x_i is considered a random variable.

R4. Dependence between observations. Functional independence is generated by the following theorem: “If two random variables X and Y are independent and $g, h: \mathbb{R} \rightarrow \mathbb{R}$, then $g(X)$ and $h(Y)$ are also independent” (Grimmitt and Stirzaker, 2001, p.49; Theorem 6A in Parzen, 1960, p.295; Theorem 2 in Ash, 1970, p.84). In SST, the sampled sites x_1 and x_2

Table 4
SAC in population, sampling, and sampled data.

Sampled data		Sampling interval	
		Within SAC range	Beyond SAC range
Population \mathfrak{R}	i.i.d. Spatial autocorrelation (SAC)	i.i.d. SAC	i.i.d. i.i.d.

Table 5
Efficiency of the spatial statistic trinity.

Population \mathfrak{R}	Sampling \mathfrak{S} and Estimator Ψ			
	Random	Stratified	Kriging	Newton
i.i.d. (neither SSH nor SAC)	Best	N/A	N/A	N/A
Spatial stratified heterogeneity (SSH)(tested by SSH q statistic)	Less efficient	BLUE	N/A	N/A
Spatial autocorrelation (SAC) (tested by Moran’s I or semivariogram)	Less efficient	Less efficient	BLUE	N/A
Deterministic (a bouncing ball)	Less efficient	Less efficient	Less efficient	Perfect

Note. “Less efficient” means too many sample units for a given MSE or a big MSE for a given sample size. BLUE denotes the best linear unbiased estimate; Newton represents a genetic determinant process and estimators; N/A stands for “not applicable”; “Deterministic” describes a generic deterministic population.

are two outcomes of an indicator random variable $I(x)$. Therefore, application of the function independence theorem for spatial sampling is incorrect, and independence between two spatial locations cannot be created by selecting the two locations independently. This contrasts with the DMF, where it is asserted that if x_1 and x_2 are random locations selected independently from each other, then the variables $z(x_1)$ and $z(x_2)$ are also independent, no matter how close the two locations are. This definition falls short because two randomly selected locations are not two random variables; thus, the theorem of function independence is not applicable to this case.

R5. Independence. In SST, samples are independent if the population is i.i.d. or the sampling interval is beyond the SAC range of the population. In truth, the properties of a population rather than the sample data determine the choice of methods for spatial statistics and

Table 6
Notations and concepts.

	DMF	Statistics (Grimmitt and Stirzaker, 2001)
Independence between two variables	If x_1 and x_2 are random locations selected independently, then they are independent, where the underline denotes random variables.	p13: Two random variables X_1 and X_2 are independent if $p(X_1, X_2) = p(X_1)p(X_2)$, where p denotes probability (p13).
Independence between two functions	If x_1 and x_2 are random locations selected independently from each other, then the variables $z(x_1)$ and $z(x_2)$ are also independent.	p49: If two random variables X and Y are independent and $g, h: \mathbb{R} \rightarrow \mathbb{R}$, then $g(X)$ and $h(Y)$ are also independent, where \mathbb{R} denotes the set of real numbers (Theorem 6A in Parzen, 1960, p295; Theorem 2 in Ash, 1970, p84).
Random variable	Two spatial sites x_1 and x_2 are two random variables.	A rv X is a process of assigning a number $X(\zeta)$ to every outcome ζ . The resulting function must satisfy the following two conditions but is otherwise arbitrary: (1) The set $\{X \leq x\}$ is an event for every x ; (2) the probability of the events $\{X = \infty\}$ and $\{X = -\infty\}$ equals 0: $p\{X = \infty\} = 0$ and $p\{X = -\infty\} = 0$. Site selection function $I(i)$ ($= 1$ if site x_i is selected, 0 otherwise) is a rv; x_i is a variable instead of a rv (Papoulis, 1991, p66).

Note. rv denotes random variable.

spatial sampling. SAC is not included in the estimation of DB (Cochran, 1978); this is no problem if the target population is i.i.d., but it may be less efficient than MB sampling (Haining, 1988; Cressie, 1993, pp.13–15) if the data are spatially autocorrelated. This is in contrast to DMF, where the ξ -independence of sample data is determined by the modeler and p -independence is determined by the sampling design.

5. Case study

According to SST, SAC in a population affects the efficiency of a sampling and the accuracy of an inference. In DB, SAC of a population is neglected or cannot be counted by a quantitative parameter (Särndal, 1978). To demonstrate the advantage of SST to DB, two case studies using a simulated dataset and a climate dataset, respectively, were carried out. In both cases, the same sampling method and estimator were adopted to draw samples and to estimate the means of several populations with different SACs. The sampling method and estimator are typical for DB and the sample size remains the same. The MSE is obtained by repeated sampling.

$$MSE = \frac{1}{T} \sum_{t=1}^T (\hat{m}_t - m)^2 \tag{7}$$

where T is the total number of sampling times for each sample size, m is the true population mean, and \hat{m}_t is the estimation of m with the sample of time t , using a DB estimator. To make the sampling errors of different populations comparable, the normalized MSE, *aveMSE*, is used:

$$aveMSE = \frac{\sqrt{MSE}}{m} \tag{8}$$

If *aveMSE* varies with the SAC of a population, SST is a reasonable framework. In both cases, the finding is consistent with different sample sizes.

5.1. Simulated dataset

Five populations with sizes of 50×50 pixels, whose values represent the Pb concentration in soil (ppm), were generated using sequential Gaussian simulation (Remy et al., 2009). Simple kriging with a spherical variogram was applied with the same sill and nugget but at different ranges, as shown in Table 7. According to the variogram, the degree of SAC in the five populations is Range40 > Range20 > Range10 > Range5 > Range1 (see Table 7 and Fig. 8).

Each of the five populations was divided into four square strata to be comparable to Brus and Guijter’s (1997) paper; samples with different sizes (16, 24, 32, 40, 48, 56, and 64) were drawn using simple random sampling (DB). For each sample size, 500 instances were randomly drawn to compute the *aveMSE* via Equation (8).

The results are shown in Fig. 9: the *aveMSE* values of simulated populations for all sample sizes take exactly the opposite order (Range40 < Range20 < Range10 < Range5 < Range1) to that of the SAC for all sample sizes. This means that the larger the SAC of the population, the smaller the variance inferred from samples drawn by simple random sampling, given the same sample size; the SAC in a population is one of the determinants of the variance of DB, and its impact increases with reductions in sample size.

Table 7
Parameters for five simulated populations.

Population name	Model	Range (m)	Sill	Nugget
Range40	Spherical	40	5.1	0.1
Range20	Spherical	20	5.1	0.1
Range10	Spherical	10	5.1	0.1
Range5	Spherical	5	5.1	0.1
Range1	Spherical	1	5.1	0.1

5.2. Climate dataset

The area-aggregated meteorological data of 363 cities from the summer of 2015 (June 1st–August 31st, 92 days), obtained from the “China Meteorological Data Sharing Service System” (data.cma.cn), were used as populations. Four meteorological indicators—that is, the daily mean temperature (TEM), daily mean relative humidity (RHU), sum of sunshine duration (SSD), and max wind speed (WIND)—have different SACs and represent four populations. Their SACs (r) are calculated using the following equation:

$$r = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \frac{Cov(V_i, V_j)}{\sqrt{Var(V_i)Var(V_j)}}, \tag{9}$$

where n is the number of cities, $Cov(V_i, V_j)$ is the covariance function of a meteorological factor between city i and city j , and $Var(V_i)$ and $Var(V_j)$ are variance functions of a meteorological factor of city i and city j . For TEM, the covariance is calculated using the following equation:

$$Cov(TEM_i, TEM_j) = \frac{\sum_{t=1}^T (TEM_{it} - \overline{TEM}_i)(TEM_{jt} - \overline{TEM}_j)}{T - 1}, \tag{10}$$

where T is the days of temperature data, TEM_{it} represents the daily mean temperature of city i on day t , and \overline{TEM}_i is the mean temperature of city i on all T days. The variance can be calculated using the following equation:

$$Var(TEM_i) = \frac{\sum_{t=1}^T (TEM_{it} - \overline{TEM}_i)^2}{T - 1}. \tag{11}$$

The SAC coefficients of different meteorological factors (populations) are listed in Table 8, and TEM > RHU > WIN > SSD.

For each meteorological indicator, the observed values of all the cities on each day are composed of a population. The sample sizes are 16, 24, 32, 40, 48, 56, and 64. Five hundred samples were randomly drawn for each sample size. The *aveMSE* of the meteorological factors of each day were calculated, and the results are plotted in Fig. 10. Clearly, the order of the *aveMSE* of TEM < RHU < Win < SSD is opposite that of the SACs (Table 8) for all sample sizes. That is, the larger the SAC of the population, the smaller the variance inferred from the samples drawn by simple random sampling; the SAC in a population is one determinant of the variance of DB, and its effect increases with reductions in the sample size.

Besides the above two case studies, Liu et al. (2018) tested the SSH and SAC of rodent density in a study area, and they found that the former is significant and the latter is weak. Then, both the SSH-based Sandwich estimator (Wang et al., 2013a) and SAC-based kriging were applied to the same sample to map the population, respectively. The Sandwich map has a smaller absolute error than the kriging map, as expected by SST.

6. Discussion and conclusion

According to the gold standard of the performance of an estimator or predictor, either MB or DB alone is incomplete. In principle, DB is applicable, but it does not guarantee optimization in all cases because it neglects the mechanism generating the population. When a population

Table 8
Spatial autocorrelation coefficients of different meteorological indicators (populations).

Meteorological indicators	Spatial autocorrelation coefficients
TEM	0.180
RHU	0.099
WIN	0.087
SSD	0.068

is not i.i.d. and prior knowledge about the spatial data process (super-population or population) is available, there are more efficient strategies for spatial sampling and statistics. MB neglects the sampling design, so it fails to account for its resulting bias; moreover, from a technical perspective, different configurations of sampling sites will produce different covariance matrices, thus generating different estimation results.

Sampling is often incorrectly neglected in spatial statistics of environmental problems. Statistics, originating from gambling, usually assumes an RRS, which allows the mathematical expectation to be determined. This assumption, however, is often far from reality in the context of spatial data, where sampling is usually performed once and non-randomly, and the accuracy of statistical inference varies with the adopted sampling strategy. The accuracy of applying spatial statistics using field data depends on the sampling, including its pattern and density (Brevik et al., 2016).

Sampling efficiency and the precision of inference are determined by the SST, which provides a simple and clear sampling and inference decision tree with the following characteristics (Table 4):

- (1) If the population is deterministic, a few sample units plus a deterministic estimator are sufficient for recovering a complete picture of the population. In this case, DB is less efficient;
- (2) If the population is SAC or SSH or both, sampling and inference should be based upon SAC or SSH or both, regardless of whether the sampled data are SAC or SSH;
- (3) If the population is i.i.d., then the data must be i.i.d. In this case, MD is useless;
- (4) A DB sampling strategy will be fine if the population is i.i.d., but it will lose efficiency if the population is SSH (Equation 5.28 in Cochran, 1977, p.99; Fig. 1 in Wang et al., 2013) or SAC (Haining, 2003, p.118) $V_{opt} \leq V_{prop} \leq V_{ran}$ (Eq. 5.28 in Cochran, 1977, p.99); and
- (5) When SAC is present, DB is less efficient than kriging; in the absence of SAC, kriging fails to work. Sandwich interpolation (Wang et al., 2013a) works if the population is SSH.

For SST, we distinguish the population and sample when mentioning data. Kriging is applicable if the population is SAC—even if the sampled data are i.i.d.—if we can obtain the variogram in other ways.

To assess, predict, and manage environmental data, this paper shows that it is important to deal carefully with the properties of spatial sampling, including how these affect the estimations of environmental properties that are important for humans being and for society at large. Therefore, spatial statistical methods should be as accurate and efficient as possible to obtain the highest accuracy of subsequent research and applications. In these cases, SST provides a useful framework to guide the choice of the proper method.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study is supported by the National Natural Science Foundation of China (Grant No. 41531179 and 42071375).

References

Angulo, J.M., Yu, H.L., Langousis, A., Kolovos, A., Wang, J.F., Madrid, A.E., Christakos, G., 2013. Spatiotemporal infectious disease modeling: a BME-SIR approach. *PLoS One* 8 (9), e72168.
 Anselin, L., 1988. *Spatial Econometrics. Methods and Models*. Kluwer, Dordrecht.
 Ash, R.B., 1970. *Basic Probability Theory*. Dover Publications, New York.

Banerjee, S., Carlin, B.O., Gelfand, A.E., 2015. *Hierarchical Modeling and Analysis for Spatial Data*, second ed. CRC Press, Florida.
 Bowley, A.L., 1906. Address to the economic and statistics section of the British association for the advancement of science. *J. Roy. Stat. Soc.* 69, 540–558.
 Branham, C.E., 1989. *Soil Survey of Pottawattamie County*. Iowa. USDA-SCS. U.S. Government Print Office, Washington, DC.
 Brevik, E.C., Calzolari, C., Miller, B.A., Pereira, P., Kabala, C., Baumgarten, A., Jordán, A., 2016. Soil mapping, classification, and pedologic modeling: history and future directions. *Geoderma* 264, 256–274.
 Brus, D.J., 2019. Sampling for digital soil mapping: a tutorial supported by R scripts. *Geoderma* 338, 464–480.
 Brus, D.J., de Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). *Geoderma* 80, 1–44.
 Chen, M., Yue, S., Lu, G.-N., Lin, H., Yang, C., Hou, T., Xiao, D., Jiang, H., 2019. Teamwork-oriented integrated modeling method for geo-problem solving. *Environ. Model. Software* 119, 111–123.
 Christakos, G., 2005. *Random Field Models in Earth Sciences*. Dover Publications, New York.
 Cochran, W.G., 1977. *Sampling Techniques*, third ed. John Wiley & Sons, New York.
 Cressie, N., 1990. The origin of Kriging. *Math. Geol.* 22 (3), 239–252.
 Cressie, N., 1993. *Statistics for Spatial Data*. Wiley, New Jersey.
 de Gruijter, J.J., ter Braak, C.J.F., 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Math. Geol.* 22, 407–415.
 de Gruijter, J., Brus, D., Bierkens, M., Knotters, M., 2006. *Sampling for Natural Resource Monitoring*. Springer, New York.
 Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. Lond.* 222, 309–368.
 Gao, B.B., Wang, J.F., Fan, H.M., Xu, K., Hu, M.-G., Chen, Z.-Y., 2015. A stratified optimization method for a multivariate marine environmental monitoring network in the Yangtze River estuary and its adjacent sea. *Int. J. Geogr. Inf. Sci.* 29, 1332–1349.
 Gao, B., Hu, M., Wang, J., Xu, C., Chen, Z., Fan, H., Ding, H., 2020. Spatial interpolation of marine environment data using P-MSN. *Int. J. Geogr. Inf. Sci.* 34 (3), 577–603.
 Goodchild, M.F., 2004. The validity and usefulness of laws in geographic information science and geography. *Ann. Assoc. Am. Geogr.* 94 (2), 300–303.
 Goovaerts, P., 1997. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, p. 187.
 Griffith, D.A., 2005. Effective geographic sample size in the presence of spatial autocorrelation. *Ann. Assoc. Am. Geogr.* 95 (4), 740–760.
 Griffith, D.A., Haining, R., Arbia, G., 1994. Heterogeneity of attribute sampling error in spatial data sets. *Geogr. Anal.* 26 (4), 300–320.
 Grimmett, G.R., Stiraker, D.R., 2001. *Probability and Random Process*, third ed. Oxford University Press, Oxford.
 Gujarati, D.N., Porter, D.C., 2009. *Basic Econometrics*, 5th Edition. McGraw-Hill Irwin, Boston.
 Haining, R., 1988. Estimating spatial means with an application to remote sensing data. *Commun. Stat.-Theor. Meth.* 17 (2), 537–597.
 Haining, R., 2003. *Spatial Data Analysis: Theory and Practice*. Cambridge University Press, London.
 Heckman, J.J., 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–161.
 Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47, 663–685.
 Isaaks, E., Srivastava, R., 1989. *Applied Geostatistics*. Oxford University Press, Oxford.
 Kiaer, A.N., 1896. Observations et expériences concernant des dénombrements représentatifs. *Bull. Int. Statist. Inst.* 9, 176–183.
 Kolovos, A., Angulo, J.M., Modis, K., Papanonopoulos, G., Wang, J.F., Christakos, G., 2013. Model-driven development of covariances for spatiotemporal environmental health assessment. *Environ. Monit. Assess.* 185 (1), 815–831.
 Krige, D.G., 1951. A Statistical Approach to Some Mine Valuations and Allied Problems at the Witwatersrand. Master's thesis. University of Witwatersrand.
 Kwan, M.P., Schwanen, T., 2018. Context and uncertainty in geography and GIScience: advances in theory, method, and practice. *Ann. Assoc. Am. Geogr.* 108 (6), 1473–1475.
 Liu, T.J., Wang, J.F., Xu, C., Ma, J.Q., Xu, C.D., Zhang, H.Y., 2018. Sandwich mapping of rodent density in Jilin Province, China. *J. Geogr. Sci.* 28 (4), 445–458.
 Matheron, G., 1963. Principles of geostatistics. *Econ. Geol.* 58, 1246–1266.
 Matheron, G., 1971. *The Theory of Regionalized Variables and its Applications*. École Nationale Supérieure des Mines de Paris, Paris.
 Minasny, B., McBratney, A.B., 2016. Digital soil mapping: a brief history and some lessons. *Geoderma* 264, 301–311.
 Mindham, D.A., Tych, W., 2019. Dynamic harmonic regression and irregular sampling: avoiding pre-processing and minimising modelling assumptions. *Environ. Model. Software* 121, 104503.
 Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. *Biometrika* 37, 17–23.
 Neyman, J., 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. Roy. Stat. Soc.* 109, 558–606.
 Papoulis, A., 1991. *Probability, Random Variables, and Stochastic Process*, third ed. McGraw-Hill, New York.
 Parzen, E., 1960. *Modern Probability Theory and its Applications*. John Wiley & Sons, New York.
 Remy, N., Boucher, A., Wu, J., 2009. *Applied Geostatistics with SGeMS: A User's Guide*. Cambridge University Press, New York.
 Ripley, B.D., 1981. *Spatial Statistics*. John Wiley Sons, New York.

- Rodriguez-Iturbe, I., Mejia, J.M., 1974. The design of rainfall networks in time and space. *Water Resour. Res.* 10, 713–728.
- Särndal, C.E., 1978. Design-based and model-based inference in survey sampling. *Scand. J. Stat.* 5, 27–52.
- Särndal, C.E., Swensson, B., Wretman, J., 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York.
- Stein, A., Hoogerwerf, M., Bouma, J., 1988. Use of soil map delineations to improve (co) kriging of point data on moisture deficits. *Geoderma* 43, 163–177.
- Sterba, S.K., 2009. Alternative model-based and design-based frameworks for inference from samples to populations: from polarization to integration. *Multivariate Behav. Res.* 44, 711–740.
- Sui, D.Z., 2004. Tobler's first law of geography: a big idea for a small world? *Ann. Assoc. Am. Geogr.* 94 (2), 269–277.
- Tober, W., 1970. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* 46 (2), 234–240.
- Wang, H., Xuan, Y., 2020. SRS-GDA: a spatial random sampling toolbox for grid-based hydro-climatic data analysis in environmental change studies. *Environ. Model. Software* 124, 104598.
- Wang, J.F., McMichael, A.J., Meng, B., Becker, N.G., Han, W.G., Glass, K., Wu, J.L., Liu, X.H., Liu, J.Y., Li, X.W., Zheng, X.Y., 2006. Spatial dynamics of an epidemic of severe acute respiratory syndrome in an urban area. *Bull. World Health Organ.* 84, 965–968.
- Wang, J.F., Christakos, G., Hu, M.G., 2009. Modeling spatial means of surfaces with stratified nonhomogeneity. *Ieee. T. Geosci. Remote.* 47 (12), 4167–4174.
- Wang, J.F., Reis, B.Y., Hu, M.G., Christakos, G., Yang, W.Z., Sun, Q., Li, Z.J., Li, X.Z., Lai, S.J., Chen, H.Y., Wang, D.C., 2011. Area disease estimation based on sentinel hospital records. *PLoS One* 6 (8), e23428.
- Wang, J.F., Hu, M.G., Xu, C.D., Christakos, G., Zhao, Y., 2012. Estimation of citywide air pollution in Beijing. *PLoS One* 8 (1), e53400.
- Wang, J.F., Haining, R., Liu, T.J., Li, L.F., Jiang, C.S., 2013a. Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface. *Environ. Plann.* 45 (10), 2515–2534.
- Wang, J.F., Jiang, C.S., Hu, M.G., Cao, Z.D., Guo, Y.S., Li, L.F., Meng, B., 2013b. Design based spatial sampling: theory and implementation. *Environ. Model. Software* 40, 280–288.
- Wang, J.F., Zhang, T.L., Fu, B.J., 2016. A measure of spatial stratified heterogeneity. *Ecol. Indicat.* 67, 250–256.
- Whitney, M., 1909. *Soils of the United States*. U.S. Department of Agriculture Bureau of Soils Bulletin 55. U.S. Government Print Office, Washington, DC.