Spatial Statistics 22 (2017) 68–88

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Spatial subsemble estimator for large geostatistical data

Márcia H. Barbian [a,*], Renato M. Assunção [b]

[a] Departamento de Estatística, Universidade Federal do Rio Grande do Sul - Porto Alegre, Brazil
[b] Departamento de Ciência da Computação, Universidade Federal de Minas Gerais - Belo Horizonte, Brazil

ARTICLE INFO

Article history:
Received 8 October 2016
Accepted 13 August 2017
Available online 24 August 2017

Keywords:
Large spatial data
Subsampling
U-statistics
Parallel computing
Gaussian random field
Maximum likelihood estimation

ABSTRACT

We introduce the concept of spatial subsemble, a subset ensemble estimation method useful in the analysis of large spatial random field datasets. The full dataset is sampled to give small spatially structured subsets of observations whose parameters are easily estimated; these are combined using a weighting scheme based on their cross-validation prediction ability. We show that our estimator is consistent. More importantly, we compare the spatial subsemble with competing alternatives and show that our proposed procedure is both accurate and much faster than its competitor. We illustrate the use of our method using several examples from large datasets.

1. Introduction

The computer revolution is still going on after decades. Presently, one of its marked aspects is the generation of massive amounts of data. We need to face datasets characterized by the presence of the 4 V's of big data: large volume, velocity, variety, and veracity. Mechanisms generating these data are found wherever remote sensors, satellites, and mobile devices are used, including the emerging internet of things. In particular, the size of spatial datasets has increased dramatically in the recent past with the growth of global satellite imaging, climate monitoring, and the remote recording of meteorological and air quality measurements. NOAA (the National Oceanic & Atmospheric Administration, part of the US Department of Commerce) has a website where terabytes of data are generated to represent the Earth's climate, atmospheric and meteorological states.

* Corresponding author.
E-mail address: helena.barbian@ufrgs.br (M.H. Barbian).

http://dx.doi.org/10.1016/j.spasta.2017.08.004
2211-6753/

The main consequence of this growth for statisticians working in spatial statistics is the need to deal with numerical and computational difficulties brought about by the massive amount of data to be analyzed, since many methods typically crash or take too long to be useful. Consider, for example, the exact computation of the likelihood assuming a Gaussian geostatistical model, with n stations located in an irregular way on a map, generally requires $O(n^3)$ numerical operations and $O(n^2)$ memory space (Stein, 2008). These numbers scale up quickly. When $n = 1000$, most spatial statistics software packages solve the problem easily, but when $n = 50\,000$, the problem becomes a severe challenge for most machines and softwares.

Several statisticians are actively looking for improved methods to analyze large spatial datasets, both for cases where the covariance function is stationary, and where it is non-stationary. In the case of stationary processes, there are two main lines of approach. The first adopts a Bayesian viewpoint with work concentrated into two main categories: using a latent process with reduced dimension (Banerjee et al., 2008; Finley et al., 2009), and using a Markov random field to approximate the Gaussian field (Lindgren et al., 2011; Rue and Tjelmeland, 2002). The second approach has more variations. One is to taper the covariance function by setting the covariance between distant stations equal to zero (Kaufman et al., 2008; Furrer et al., 2006). Another possibility is to truncate the spectral representation to zero (Fuentes, 2007). Stein et al. (2004) and Vecchia (1988) used a composite likelihood function while Sun and Stein (2016) work with the score function, with its inverse covariance matrix approximated by a sparse matrix, and Castrillón-Candás et al. (2016) use multilevel set of contrasts. All these methods ignore some aspect of the full model in order to reduce the numerical complexity. Most of them differ by selecting different aspects to achieve a simplified likelihood function.

The non-stationary large spatial datasets case has a smaller volume of published research. Recently, results have been published by Datta et al. (2016), Katzfuss (2016), Konomi et al. (2014), Katzfuss (2013) and Sang et al. (2011).

Analysis of non-spatial big data problems are to be addressed by statisticians using subsampling techniques (Kleiner et al., 2014; Sapp et al., 2014) and divide and conquer techniques (Guha et al., 2012; Chen and Xie, 2014), which can significantly reduce the dimension of the problem, hence they can alleviate the computational demand; a review of these techniques can be found in Schifano et al. (2016) and Bühlmann et al. (2016). In the context of spatial statistics, Liang et al. (2013) developed a method for large geostatistical datasets using a resampling-based, in which small subsamples are sequentially selected and model parameter estimates are updated within the framework of stochastic approximation of Robbins and Monro (1951) and Andrieu et al. (2005).

Liang et al. (2013) provide a fast and consistent estimator which, hence it is appropriate for large datasets. However, their method has drawbacks: it has a sequential structure which prevents it from being parallelized, and it is necessary to check the stochastic convergence of the algorithm.

In this paper, we propose a new method that is simple to apply, computationally fast and requires little memory space. It allows for the calculation of confidence intervals and it has good theoretical properties for both the infill asymptotic approach as well as for the increasing domain asymptotic approach. The method is based on subsampling small spatially structured subsets of observations. In each subsample, we fit the parameters with the preferred method and combine them using a validation subset. The two main advantages of our method are: first, its simplicity, making it very easy to use; and second, its speed, since it can be parallelized. To show these advantages, we compare our spatial subsemble algorithm with resampling-based stochastic approximation (RSA) by Liang et al. (2013), and MLE (Maximum Likelihood Estimation), showing that the proposed method is accurate and substantially faster than the other methods. We illustrate our method using a large NOAA dataset of precipitation over the United States.

2. The spatial subsemble estimator

2.1. The geostatistical Gaussian model

Suppose the vector $\mathcal{Y} \equiv (Y(s_1), Y(s_2), \ldots, Y(s_n))^T$ are observed values of a random process $\{Y(s) : s \in D \subset \mathbb{R}^2\}$, where the spatial index s varies continuously across the region D. Let the

geostatistical Gaussian model be:

$$Y(s_i) = \mu(s_i) + X(s_i) + \epsilon(s_i), \tag{1}$$

where $\mu(s_i) = \mathbb{E}[Y(s_i)]$ and $(\epsilon(s_1), \epsilon(s_2), \ldots, \epsilon(s_n))$ are independent normal, identically distributed random variables with mean zero and variance τ^2, the nugget effect. The random variables $(X(s_1), X(s_2), \ldots, X(s_n))$ come from a Gaussian process $\{X(s) : s \in D \subset \mathbb{R}^2\}$ with $\mathbb{E}[X(s) = 0]$ and covariance matrix equal to $\sigma^2\mathbf{R}$. The $n \times n$ matrix $\mathbf{R}$ has elements $R_{ij} = \rho(\|s_i - s_j\|; \boldsymbol{\lambda})$ representing the correlation between $X(s_i)$ and $X(s_j)$. The Euclidean distance between s_i and s_j is $\|s_i - s_j\|$ and $\boldsymbol{\lambda}$ is the set of correlation function parameters.

In this model, $\mathbf{Y}$ has a multivariate normal distribution with mean vector $\mu\mathbf{1}$ and covariance matrix $\boldsymbol{\Sigma} = \sigma^2\mathbf{R} + \tau^2\mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{1}$ is an n-dimensional vector composed of 1's.

We can generalize the above model further by allowing for a non-constant mean represented by

$$\mu(s_i) = \beta_0 + \sum_{t=1}^{p} \beta_t \mathbf{C}_t(s_i), \tag{2}$$

where $\mathbf{C}_t(s_i)$ denotes the tth explanatory variable evaluated at s_i and β_t is the corresponding regression coefficient. For details on this and other models see Cressie (2015).

Let $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_p, \tau^2, \sigma^2, \boldsymbol{\lambda})$ represent the set of parameters in this geostatistical model. Then the likelihood function is given by:

$$L(\boldsymbol{\theta}, \mathbf{y}) = -\frac{n\log(2\pi)}{2} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}). \tag{3}$$

2.2. The spatial subsemble idea

We propose a divide-and-conquer strategy (Guha et al., 2012) adapted for the spatial context. The idea is intuitively simple and is explained in Fig. 1. The points shown in the square form the large collection of n sampling stations which make the likelihood computation infeasible. We subsample $m = jk$ spatially structured stations; this subsample is composed by j sets of k spatially-close stations. In Fig. 1, this is illustrated by the $j = 4$ sets of $k = 12$ stations which are orange colored. This small subsample is used to estimate the parameter $\boldsymbol{\theta}$ producing $\hat{\boldsymbol{\theta}}$ by a method such as maximum likelihood. A mixing weight w is associated with this subsample based on validation data shown as blue stations. Using $\hat{\boldsymbol{\theta}}$ and the observed $Y(s)$ values at the blue stations we predict the values at the red stations. The mean squared prediction error at these red stations gives a measure of quality for $\hat{\boldsymbol{\theta}}$ and the mixing weight is inversely proportional to this measure.

To compensate for the small subsample size, the spatial subsemble estimation procedure is repeated independently B times. It combines the resulting B estimates of $\boldsymbol{\theta}$ using a weighted mean with the mixing weights w. We provide next a formal description of our method.

2.3. The definition of the spatial subsemble estimator

Let $\mathbf{Z}(\mathbf{s}) = (Y(s_1^*), Y(s_2^*), \ldots, Y(s_m^*))^T$ be a subsample from the dataset $\mathcal{Y} = \{Y(s_1), Y(s_2), \ldots, Y(s_n)\}$ with m being much less than n. Given $\mathbf{s} = (s_1^*, s_2^*, \ldots, s_m^*)$, the vector $\mathbf{Z}$ has the same underlying model as $\mathbf{Y}$:

$$\mathbf{Z}|\mathbf{S} \sim N_m(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \tag{4}$$

where $\boldsymbol{\Sigma}_z = \sigma^2\mathbf{R}_z + \tau^2\mathbf{I}_m$, $\mathbf{R}_z$ is the $m \times m$ correlation matrix of $\mathbf{Z}(\mathbf{s})$ and $\boldsymbol{\mu}_z = (\mu(s_1^*), \mu(s_2^*), \ldots, \mu(s_m^*))^T$ is its expected value.

Liang et al. (2013) consider a U-statistic approximation for the Kullback–Leibler divergence and its minimization leads to their estimating equation

$$\binom{n}{m}^{-1} \sum_{i=1}^{\binom{n}{m}} \frac{\partial \log f_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathbf{s}_i)}{\partial\boldsymbol{\theta}} = \mathbf{0},$$

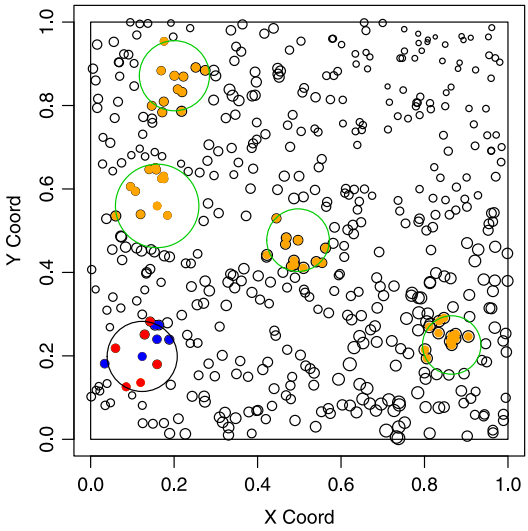

Fig. 1. Example of spatial subsemble: orange points represent subsample of size $m = 48$ with $j = 4$ and $k = 12$, blue points are the validation subset and red points are the prediction subset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where $f_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathbf{s}_i)$ is the joint multivariate normal density of model (4). This sum considers all subsamples of size m from the n stations. Rather than using all subsamples, Liang et al. (2013) use the Robins–Monro stochastic approximation algorithm; they sequentially subsample sets of m stations and update the estimate $\hat{\boldsymbol{\theta}}$ iteratively. This algorithm needs tuning parameters, one of them is used to define the acceptance rate of the algorithm. The other is used in the estimation equations, indicating the difference between updates.

The selection of m stations does not have any spatial structure, randomly choosing among the n possible stations. This carries less information about the spatial dependence than adjacent stations, especially for the dependence among nearby locations.

Instead, we propose to independently select a spatially structured subsample of size $m = jkB$ times and to combine them with weights associated with their prediction quality. Our subsample should reflect the large field in a reduced scale. Given the ith subsample $\mathbf{Z}(\mathbf{s}_i) = (Y(s^*_{i1}), Y(s^*_{i2}), \ldots, Y(s^*_{im}))^T$ at the locations $\mathbf{s}_i = (s^*_{i1}, \ldots, s^*_{im})$, we obtain the ith maximum likelihood estimator

$$\hat{\boldsymbol{\theta}}_i = \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{z}_i|\mathbf{s}_i).$$

We want to be able to estimate $\boldsymbol{\theta}$ as well as possible given the small subsample size m. For this, we require a combination of both stations near to each other, and stations far apart. The k stations located next to each other provide information about the short-range correlation $\rho(\cdot\,; \boldsymbol{\lambda})$. The stations far apart allow us to estimate the range parameter. Therefore, we randomly select j centers in the field and their k nearest stations.

We also select v other sets of k neighboring stations to obtain a prediction quality for the subsample estimates. These stations are randomly split into two subsets: a validation subset $\mathbf{Z}^v$ and a prediction subset $\mathbf{Z}^p$. We use the validation subset $\mathbf{Z}^v$ to build the best linear unbiased predictor (BLUP) for the locations in the prediction subset $\mathbf{Z}^p$. The BLUP estimator (Diggle and Ribeiro Jr, 2007) is given by

$$\widehat{\mathbf{Z}^p} = \hat{\boldsymbol{\mu}}_{z^p} + \mathbf{r}'\boldsymbol{\Sigma}^{-1}_{z^v}\left(\mathbf{Z}^v - \hat{\boldsymbol{\mu}}_{z^v}\right),$$

where $\mathbf{r}$ is the covariance between $\mathbf{Z}^v$ and $\mathbf{Z}^p$ as estimated by $\hat{\boldsymbol{\theta}}$. The prediction quality weight is given by

$$w^{-1} = \|\mathbf{Z}^p - \widehat{\mathbf{Z}^p}\|^2.$$

Since this is repeated B times independently, let $i = 1, \ldots, B$ index the subsample and $\mathbf{s}_i$ the stations composing the i-subsample. We denote the estimator of the parameters based on this ith subsample by $\hat{\boldsymbol{\theta}}_i$. The precision quality is w_i and our spatial subsemble estimator is given by

$$\hat{\boldsymbol{\theta}} = \frac{\sum_i w_i \hat{\boldsymbol{\theta}}_i}{\sum_i w_i}. \tag{5}$$

A simpler alternative estimator is given by

$$\tilde{\boldsymbol{\theta}} = \frac{\sum_i \hat{\boldsymbol{\theta}}_i}{B}. \tag{6}$$

Using U-statistics and subsampling theory, we find the theoretical properties of $\tilde{\boldsymbol{\theta}}$ in Section 5. Equivalent properties for $\hat{\boldsymbol{\theta}}$ are much harder to prove due to the presence of the stochastic weights in (5). However, in the simulated studies we conducted, they had a performance similar to that of $\tilde{\boldsymbol{\theta}}$ (see Section 3).

Our estimators are examples of the general strategy known as *divide-and-conquer*. It is based on the idea of breaking down the problem into several pieces in each of which estimation task is straightforward. Next, we combine the results from these small pieces producing a single final estimate. One of the main advantages of this strategy is the possibility of processing each piece separately and independently, making it possible to use multithreading programming (Tanenbaum, 2009). Each thread can be sent and separately estimated by a CPU core; the greater the number of cores, the faster the computation. Another advantage of our approach is that it provides the calculation of confidence intervals.

To obtain an approximation for the variance of our estimators, consider conditioning on $\mathcal{I} = \{X(s) : s \in D\} \cup \{\epsilon(s_1), \ldots, \epsilon(s_n)\}$, the random Gaussian field and the nugget random terms. Then

$$\mathbb{V}(\hat{\boldsymbol{\theta}}) = \mathbb{V}\left(\mathbb{E}(\hat{\boldsymbol{\theta}}|\mathcal{I})\right) + \mathbb{E}\left(\mathbb{V}(\hat{\boldsymbol{\theta}}|\mathcal{I})\right). \tag{7}$$

As the B samples are selected independently given $\mathcal{I}$, we can estimate

$$\mathbb{E}\left(\mathbb{V}(\hat{\boldsymbol{\theta}}|\mathcal{I})\right) = \mathbb{E}\left(\frac{\sum_i w_i^2 \mathbb{V}(\hat{\boldsymbol{\theta}}_i|\mathcal{I})}{\left(\sum_i w_i\right)^2}\right) \widehat{=} \frac{\sum_{i=1}^{B} w_i^2 \mathbb{V}(\hat{\boldsymbol{\theta}}_i)}{\left(\sum_i w_i\right)^2} \tag{8}$$

where $\mathbb{V}(\hat{\boldsymbol{\theta}}_i)$ is an approximation for the covariance matrix of $\hat{\boldsymbol{\theta}}_i$, given by Mardia and Marshall (1984) as the inverse of

$$IF(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{Y}) = \text{diag}(G_{\boldsymbol{\beta}}, G_{\boldsymbol{\gamma}}),$$

where $G_{\boldsymbol{\beta}} = \mathbf{C}^T \boldsymbol{\Sigma} \mathbf{C}$, $\mathbf{C}$ is a regressor matrix, $\boldsymbol{\gamma} = \{\boldsymbol{\lambda}, \sigma^2, \tau^2\}$ and the (i, j)th element of $G_{\boldsymbol{\gamma}}$ is given by

$$G_{\boldsymbol{\gamma}}(i,j) = \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \gamma_i}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \gamma_j}\right) = \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \gamma_i}\boldsymbol{\Sigma}\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \gamma_j}\right).$$

In the applications we are considering, it seems reasonable to expect that the first term on the right hand side of (7) is close to zero or, at least, much smaller than the second term. The reason is that $\mathbb{E}(\hat{\boldsymbol{\theta}}|\mathcal{I})$ is likely to be close to its unconditional expectation $\mathbb{E}(\hat{\boldsymbol{\theta}})$ if B is large and the j centers are separated apart by distances larger than the spatial correlation range. As geographically distant stations are approximately independent random variables, in a large B number, we should have $\mathbb{E}(\hat{\boldsymbol{\theta}}|\mathcal{I}) \approx \mathbb{E}(\hat{\boldsymbol{\theta}})$. Hence, the variance of $\mathbb{E}(\hat{\boldsymbol{\theta}}|\mathcal{I})$ should be close to zero.

A similar argument involving $\tilde{\boldsymbol{\theta}}$ leads to an approximation for the variance of our estimators given by

$$\mathbb{V}(\hat{\boldsymbol{\theta}}) = \frac{\sum_i w_i^2 \mathbb{V}(\hat{\boldsymbol{\theta}}_i)}{\left(\sum_i w_i\right)^2} \tag{9}$$

and

$$\mathbb{V}(\tilde{\boldsymbol{\theta}}) = \frac{\sum_i \mathbb{V}(\hat{\boldsymbol{\theta}}_i)}{B}. \tag{10}$$

3. Performance under simulation

We use Monte Carlo simulation to investigate the performance of our estimator. We analyze the impact of different choices for the number of replications B, the subsample size m, and the type of spatial subsampling undertaken. We compare the results with the maximum likelihood estimator (MLE) based on the complete dataset and with the estimator RSA proposed by Liang et al. (2013). The simulation scenarios used in this paper are very similar to those used by Liang et al. (2013).

As our estimator can use parallel programming, much processing time can be saved, with savings increasing along with the increased number of processors. To evaluate this behavior, we used two hardware configurations to calculate the estimates: one with 4 cores and another with 24 cores. All results were calculated on the SGI Altix processor cluster, operating under Novell SUSE Linux Enterprise Server and a connection unit with 64 processing units. Each unit has 64 GB of RAM and two processors dodecacore AMD Opteron (32 units with the 6176 SE model 2.3 GHz and the other 32 units with the model of 2.9 GHz frequency). To access the cloud it was possible to use only one processing unit in each simulation run.

The software used to generate the data, to carry out the estimation and to analyze the results was `R`, version 3.2.3. The package `geoR` (Ribeiro and Diggle, 2001) generated the data analyzed, calculated the MLE and our proposed estimator. The RSA estimator was implemented through a package kindly provided by these authors. The package `parallel` was used to run the parallel computations.

The function available in `geoR` to calculate the MLE requires initial values for the optimization method. Several tests were carried out in which we maximize the likelihood with fixed and randomly chosen initial values. The estimate was basically the same in both situations. Hence, the rest of this paper presents only the results with random initial values.

The data are simulated from a Gaussian random field following (1) and (2), with irregularly spaced sampling stations in the region $D = [0, 100] \times [0, 100]$. The correlation function assumed is the exponential function $\rho(\|s_i - s_j\|; \phi) = \exp^{-\|s_i - s_j\|/\phi}$. There is only one covariate C_1 with normal distribution with mean 0 and standard deviation 0.5. The model has the following parameter values: $\sigma^2 = 1$, $\beta_0 = 1$ and $\beta_1 = 1$.

We consider four different scenarios:

- Scenario 1: We simulated $Y(s_i)$, $i = 1, \ldots, n$ with $n = 2000$ observations, $\phi = 25$, and $\tau^2 = 1$.
- Scenario 2: $n = 2000$ and $\phi = 25$ as above but $\tau^2 = 0$.
- Scenario 3: $n = 50000$ observations with $\phi = 25$ and $\tau^2 = 1$.
- Scenario 4: $n = 2000$, $\phi = 5$ and $\tau^2 = 1$.

Scenario 4 aims to evaluate the effect of sampling in the context of the increasing domain asymptotics when the study region tends to increase together with the number of sampling stations (see Section 5). In this situation, the MLE is consistent and asymptotically normal, making it possible to obtain asymptotic confidence intervals (CI) for the model parameters. Scenarios 1, 2, and 3 are examples of the infill asymptotics approach, when not all parameters in the Matérn class can be estimated consistently and hence the individual results for ϕ and σ^2 are not shown. Instead, we give the estimates for the ratio ϕ/σ^2, which can be consistently estimated (Zhang, 2004). The sample size in scenario 3 makes infeasible the MLE calculation with the complete dataset. Therefore, in this scenario, only the results with the RSA and subsemble estimators are reported.

In each scenario, we evaluated the performance of our estimator under three different subsampling schemes shown in Fig. 2.

- $5C$, five centers: we selected $j = 5$ central stations as well as and their $k = \frac{m}{5} - 1$ nearest stations. The selection algorithm of the subsample is done sequentially: a center point is selected together with its nearest k observations. These stations are then excluded from the selection set. This procedure is repeated with the remaining stations with j iterations.
- $1C$, one single center: a central sampling station is randomly selected as well as its $m-1$ nearest sampling stations.
- $1CD$, one center plus distant stations: we selected one station as a single center as well as its $m - 5 - 1$ nearest stations. We then add the 5 stations located the farthest from the selected center.

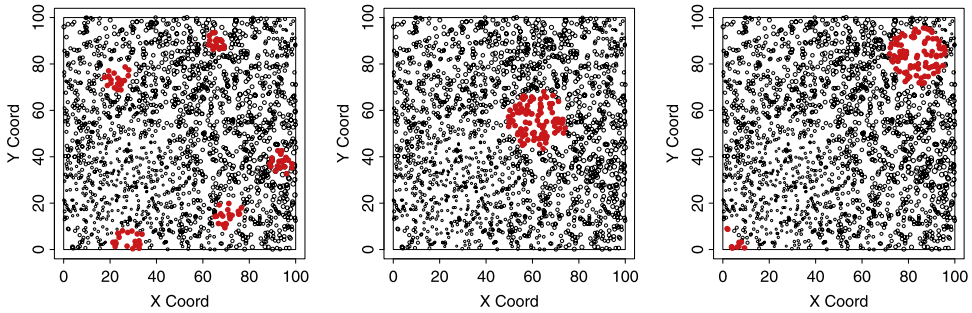

Fig. 2. Examples of selection methods for subsamples of size $m = 100$ (red points) in a Gaussian field with 2000 observations: 5 centers (left), 1 center (middle) and 1 center plus distant (right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To calculate the weights w in each subsample, 50 spatially grouped observations were drawn. The method to select these points was 1C ($j = 1$ and $k = 50 - 1$) and the selection did not include the stations used in the estimation step. Out of this subset, 25 stations were randomly assigned to the vector $\mathbf{Z}^v$, the other 25 stations forming the prediction group $\mathbf{Z}^p$.

In each case, we repeated the estimation procedure 50 times. The small number of simulations is due to the long time needed to process each one of them.

3.1. Characteristics of $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$

Before comparing $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ with the other methods, we studied their characteristics with respect to the choices for B, m and station selection method. The estimates were calculated with subsamples of size $m = \{100, 300, 500, 700\}$. The number B of subsamples was taken equal to 25 and 50. The station sampling design or scheme was the one center (1C), one center plus distant stations (1CD) and 5 centers (5C).

The results are based on the mean, standard deviation (in parentheses) and boxplots of the estimates calculated with 50 simulated datasets. Since the behavior is basically the same in all scenarios and due to lack of space, we show the detailed analysis for scenario 3, providing only a general description of the results for the other 3 scenarios.

Figs. 3 and 4 present the estimates of the parameters β_0 and β_1. All the estimates' distributions are very similar and their expected values close to the true values, independently of the stations selection scheme, B value, and estimator type. The most noticeable effect is that estimator with $B = 50$ has less variability than with $B = 25$.

For the ratio ϕ/σ^2, Fig. 5 shows that there are large differences among the station selection methods. The 1C scheme gave the worst results, with large underestimation, which decreases with the increase of m. The 5C and 1CD sampling schemes have similar and approximately unbiased results although there is a small bias when $m = 100$ with 5C. Analyzing the estimates using $B = 25$, we can see more variability than those using $B = 50$, independently of m, estimator and sampling scheme. Furthermore, in practice, there seems to be very little difference between $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$.

Concerning the estimation of τ^2, we have a systematic bias in all cases, as shown in Fig. 6. The subsample size has a clear influence on the estimates. The sampling scheme 1C induces the largest bias and variability, this being also evident for $m = 100$ and $m = 300$. The sampling scheme 5C also gave the best results, and once again, the estimates with $B = 50$ have smaller variance than $B = 25$.

As expected, the larger the subsample size m, the better the estimator, this aspect being more obvious for the covariance parameters. The different values of B have little effect on the bias and variance of the spatial subsemble estimators and when $B = 50$, they have slightly less variance than when $B = 25$. We experimented with larger values for B (results not shown here) finding that taking

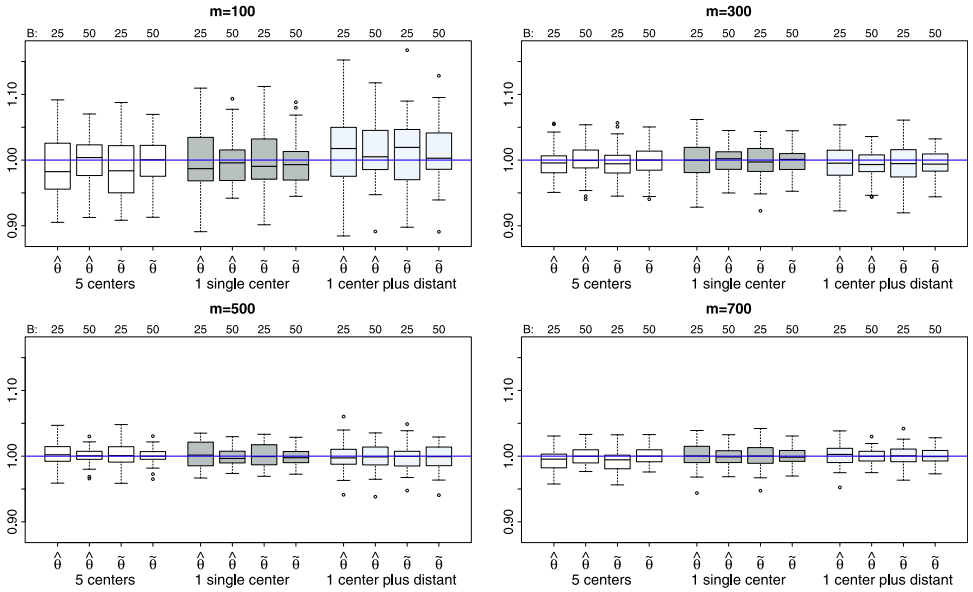

Fig. 3. Estimates of $\beta_0 = 1$ for scenario 3, given different selection algorithms, m and B. We have $m = 100$ (top left), $m = 300$ (top right), $m = 500$ (bottom left) and $m = 700$ (bottom right).

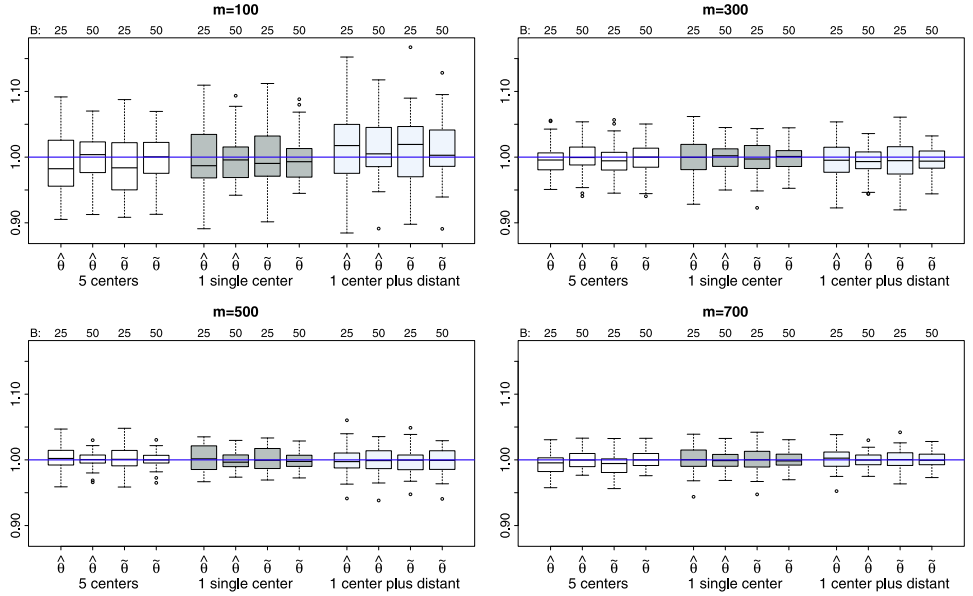

Fig. 4. Estimates of $\beta_1 = 1$ for scenario 3, given different selection algorithms, m and B. Same description as in Fig. 3.

$B > 50$ does not produce better results than taking $B = 50$. It is important to emphasize that using $B = 50$ rather than $B = 25$ takes twice as long to obtain the estimates.

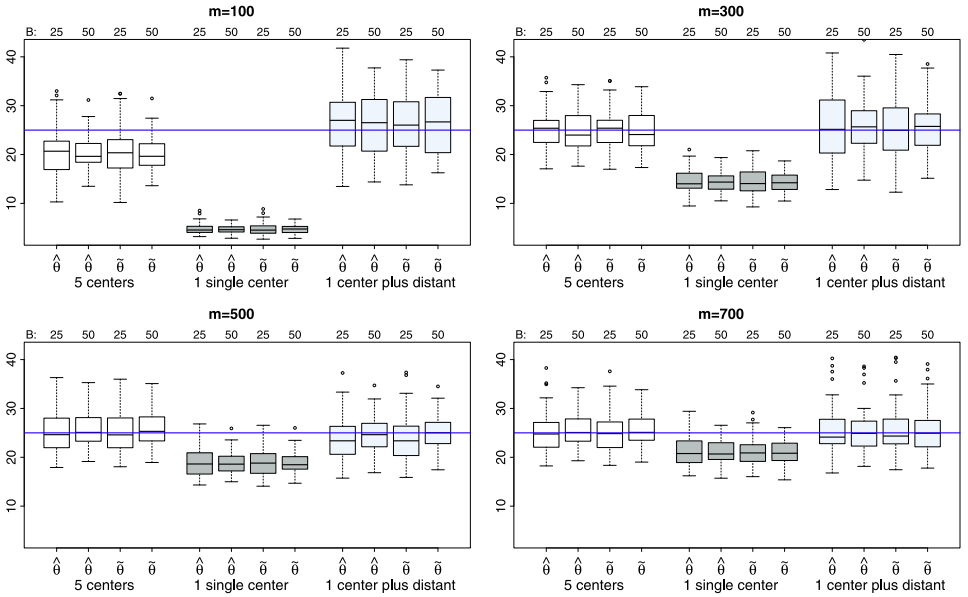

Fig. 5. Estimates of ϕ/σ^2 for scenario 3, given different selection algorithms, m and B. Same description as in Fig. 3.

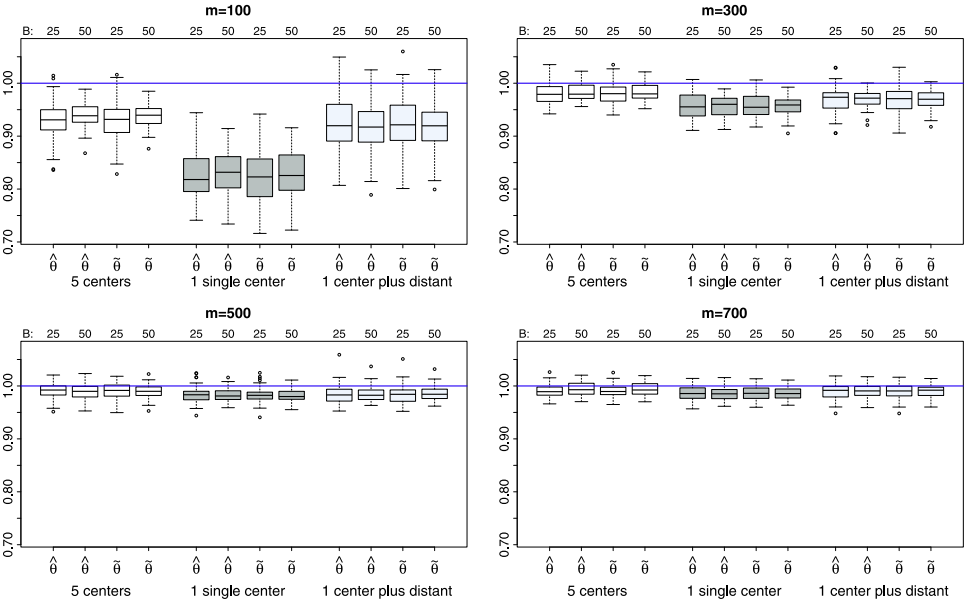

Fig. 6. Estimates of τ^2 for scenario 3, given different selection algorithms, m and B. Same description as in Fig. 3.

Estimates of β_0 and β_1 are very close to the true parameter value and this conclusion is independent of the sampling station scheme. Concerning ϕ/σ^2 and τ^2, there is a difference; the sampling schemes 1CD and 5C show a similar performance and are superior to that of 1C. In spite of that, as the subsample

Table 1
Mean processing time (minutes) of the 50 simulations for each scenario.

Method	m	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
		4 cores	24 cores	4 cores	24 cores	4 cores	24 cores	4 cores	24 cores
MLE	–	33.720	–	33.923	–	–	–	36.050	–
$\hat{\boldsymbol{\theta}}$	100	0.111	0.032	0.122	0.040	0.165	0.047	0.139	0.046
	300	1.723	0.432	2.038	0.512	2.143	0.414	1.645	0.598
	500	9.295	1.529	8.831	1.822	10.690	1.604	6.153	2.150
	700	23.731	3.831	21.596	3.969	25.827	4.119	15.609	4.549
RSA	100	0.163	–	0.182	–	0.256	–	0.164	–
	300	2.960	–	3.826	–	3.244	–	2.942	–
	500	14.753	–	10.678	–	15.525	–	14.755	–
	700	40.784	–	33.042	–	43.244	–	41.883	–

size increases, this difference among the three sampling schemes decreases with the bias tending to zero. As an overall summary, the estimators $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ have a similar performance under the situations analyzed here.

3.2. Comparison with other methods

In this subsection, we compare the subsemble estimator with RSA and with the MLE based on the full dataset. We do not cover all sampling schemes and values for B again, presenting only the results for the 5C sampling scheme and $B = 25$. As we saw, the accuracy gain with $B = 50$ was not high enough to justify twice the length of processing time. We use an upper index in the estimator symbol to indicate the subsample size. For example, RSA^{500} is the RSA estimator using a subsample equal to 500.

Besides the estimates, we measure the processing time in minutes. For the MLE and RSA estimators, we present only the estimates using 4 cores, as there is no decrease when more cores are used. For $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$, the times required to run on hardware with 4 and 24 cores are presented. As the proposed estimators were calculated jointly, the processing time is presented only for $\hat{\boldsymbol{\theta}}$. As $\tilde{\boldsymbol{\theta}}$ is simpler, not requiring the cross-validation weights, it would be faster than $\hat{\boldsymbol{\theta}}$. Table 1 shows that, for all scenarios, the time required to obtain the subsemble estimators was smaller than those required by RSA. This difference was more noticeable when we used more cores.

Scenario 1

Fig. 7 shows the results in graphical form. The β_0 and β_1 estimators provided by $\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}$, RSA and the MLE have very similar results. Irrespective of the subsample size, the mean is close to the true parameter value. Additionally, when the subsample size increases, the subsemble estimators' variance decreases and approaches the MLE variance. Concerning the ratio ϕ/σ^2 and τ^2, we see that all estimators underestimate the true parameter value; however, this bias decreases as the subsample size increases. The spatial subsemble estimator has smaller bias and variance than the RSA estimator for the ϕ/σ^2.

Scenario 2

Fig. 8 shows that, for the parameters β_0 and β_1, the estimators $\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}$ and MLE produce very similar results. For β_1, RSA has a variance substantially larger than the alternatives and has occasional outlier values. For ϕ/σ^2, the MLE, $\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}$, RSA^{100} and RSA^{300} are very similar, and all approximately unbiased. For the RSA^{500} and RSA^{700} cases, the results tend to underestimate ϕ/σ^2. For τ^2, it should be remembered that its true value is zero in this scenario, hence any positive estimate (as it must be) will necessarily overestimate it; what is most noticeable is the relatively large values obtained when using RSA.

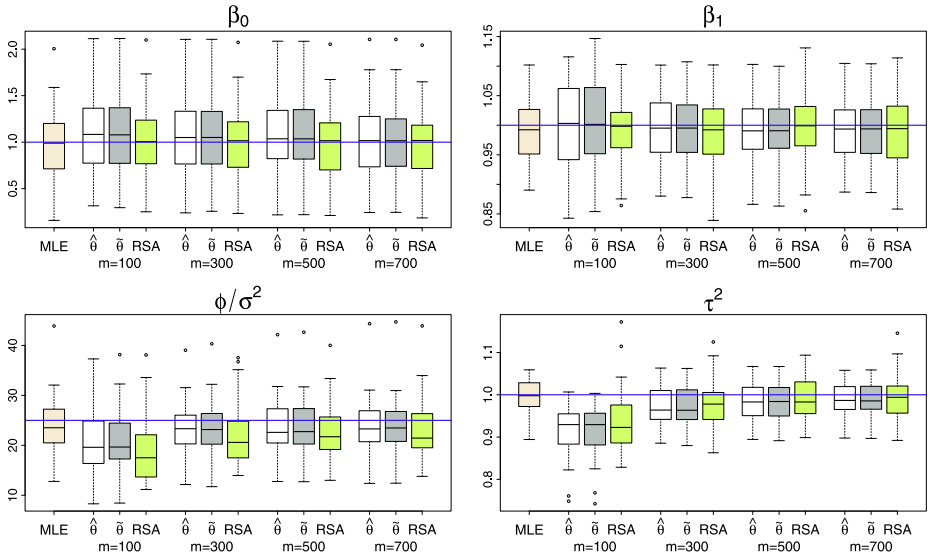

Fig. 7. Boxplot of the estimates for scenario 1. Upper left: β_0. Upper right: β_1. Lower left: ϕ/σ^2. Lower right: τ^2. Subsample sizes m are 100, 300, 500, and 700.

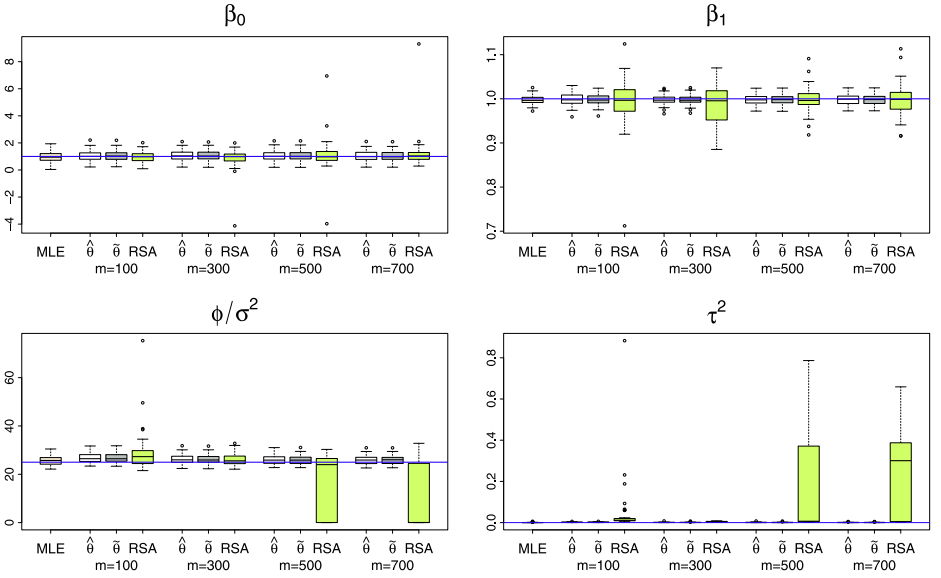

Fig. 8. Boxplot of the estimates for scenario 2. Upper left: β_0. Upper right: β_1. Lower left: ϕ/σ^2. Lower right: τ^2. Subsample sizes m are 100, 300, 500, and 700.

Scenario 3

Fig. 9 shows the results for scenario 3. This scenario differs from scenario 1 only with respect to the dataset size, with $n = 50{,}000$ here while $n = 2000$ in scenario 1. Due to the large sample size, the full dataset MLE estimates were not calculated. The results are qualitatively similar to those obtained under scenario 1, the difference being that this larger dataset situation has increased the effect already

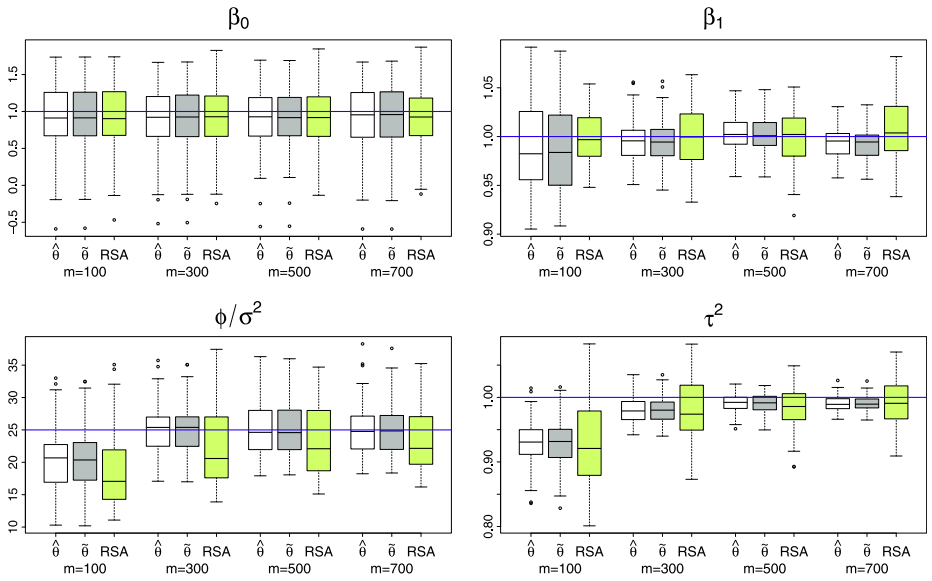

Fig. 9. Boxplot of the estimates for scenario 3. Upper left: β_0. Upper right: β_1. Lower left: ϕ/σ^2. Lower right: τ^2. Subsample sizes m are 100, 300, 500, and 700.

seen in scenario 1. There is very little difference between estimators and different subsample sizes for β_0. The parameter β_1 has RSA with a better performance than our subsemble alternative when the subsample size is small ($m = 100$, only 0.2% of the full dataset) but this is reversed for $m \geq 300$. The parameter ϕ/σ^2 is better estimated by $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$, shown most clearly for $m \geq 300$, RSA shows some persistent bias, as well as a larger variance. With respect to τ^2, all estimators underestimate the true parameter value with the bias and variance decreasing as subsample size increases. Furthermore, $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ have smaller bias and variability than the RSA estimator.

Scenario 4

Fig. 10 shows the results for scenario 4. This is a situation where $\phi = 5$ and, compared with scenario 1, correlation decreases more rapidly as the distance between stations increases. All estimators have similar behavior, except in the case of τ^2 when $m = 100$. The parameters β_0 and β_1 are estimated approximately without bias, the standard deviation of RSA and the subsemble estimators have values similar to those of the MLE. For τ^2, the results for the spatial subsemble with $m = 100$ and $m = 300$ underestimate the true parameter value. For $m \geq 500$, RSA, $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ have results close to those of the MLE, with no clear dominance of any of them.

To compare the sample distribution of the RSA and subsemble estimators with that of the MLE, we obtained kernel estimates of their densities in the case of the ϕ parameter (see Fig. 11). Note that, for $m = 700$, there are few differences among the methods. For smaller subsample sizes, the proposed estimator behaves more like the MLE estimator.

To compare the approximation quality of Eqs. (9) and (10), we obtained the CI for the MLE and the proposed estimators. Table 2 shows that the difference between the CI range of the MLE and subsemble estimators is substantial. For ϕ and $m = 100$, the MLE standard error is 12 times smaller than the $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ subsemble estimators. However, as m increases, the standard errors of the latter estimators approach those of the MLE. With respect to the interval coverage, the MLE had a smaller coverage for σ^2, and τ^2. The $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ estimators have a 100% coverage for virtually all parameters and values of m. This is due to the over estimation of the variances provided by (9) and (10). Although undesirable, this variance overestimation leads to conservative results in the sense that the true CI coverage is larger than the nominal 95%.

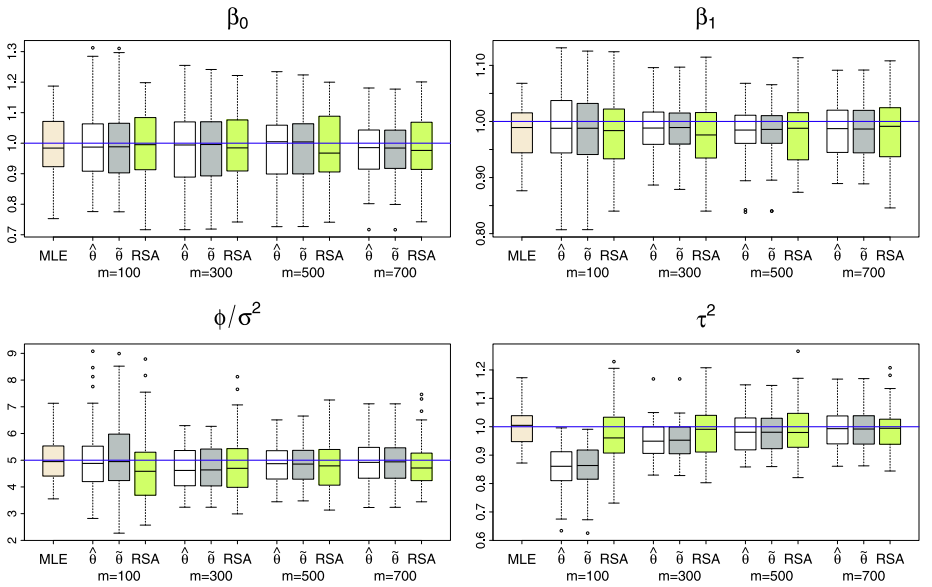

Fig. 10. Boxplot of the estimates for scenario 3. Upper left: β_0. Upper right: β_1. Lower left: ϕ/σ^2. Lower right: τ^2. Subsample sizes m are 100, 300, 500, and 700.

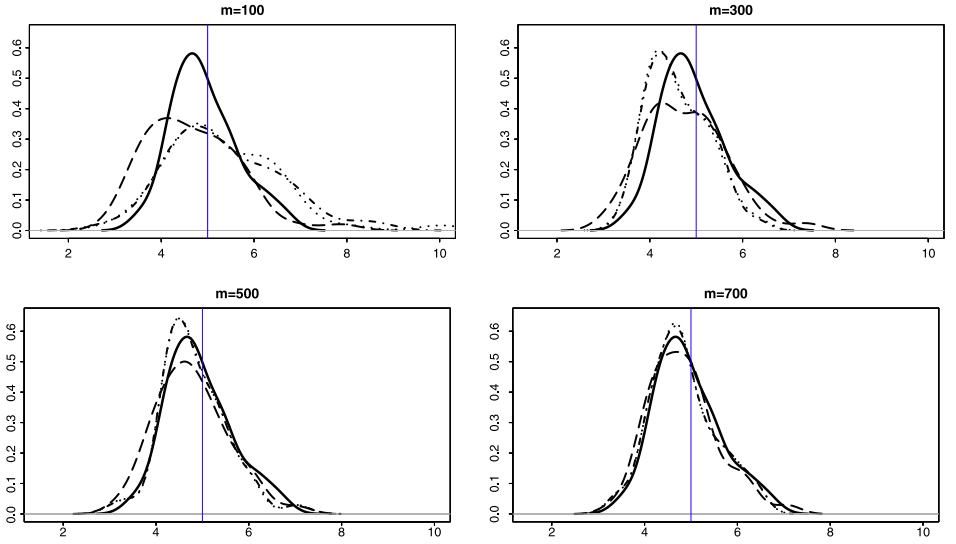

Fig. 11. Densities estimates of ϕ for scenario 4, MLE (——), $\hat{\boldsymbol{\theta}}$ (· · · · · ·), $\tilde{\boldsymbol{\theta}}$ (· — · — ·) and RSA (— — — —) for subsample $m = 100$ (top left), $m = 300$ (upper right), $m = 500$ (lower left) and $m = 700$ (lower right).

4. Application

To illustrate our method, we use publicly available data from the NOAA website, previously used by Kaufman et al. (2008), Furrer et al. (2006) and Liang et al. (2013). The data used in the analysis have 11,918 observations, which date from April, 1948. The response variable is the station-specific

Table 2
Coverage percentage (% Cov) for intervals of 95% confidence level and standard error (SE) for estimates of scenario 4.

Method	m	β_0		β_1		ϕ		σ^2		τ^2	
		% Cov	SE	% Cov	SE	% Cov	SE	% Cov	SE	% Cov	SE
MLE	–	96%	0.11	98%	0.05	96%	0.77	92%	0.11	92%	0.06
$\hat{\boldsymbol{\theta}}$	100	100%	0.29	100%	0.24	100%	8.91	100%	0.59	100%	0.45
	300	100%	0.21	100%	0.13	100%	2.19	100%	0.27	100%	0.17
	500	100%	0.18	100%	0.10	100%	1.60	100%	0.21	100%	0.12
	700	100%	0.16	100%	0.09	100%	1.33	100%	0.18	100%	0.10
$\tilde{\boldsymbol{\theta}}$	100	100%	0.29	100%	0.24	100%	8.96	100%	0.58	80%	0.12
	300	100%	0.21	100%	0.13	100%	2.19	100%	0.27	100%	0.12
	500	100%	0.18	100%	0.10	100%	1.61	100%	0.21	100%	0.12
	700	100%	0.16	100%	0.09	100%	1.33	100%	0.19	100%	0.10

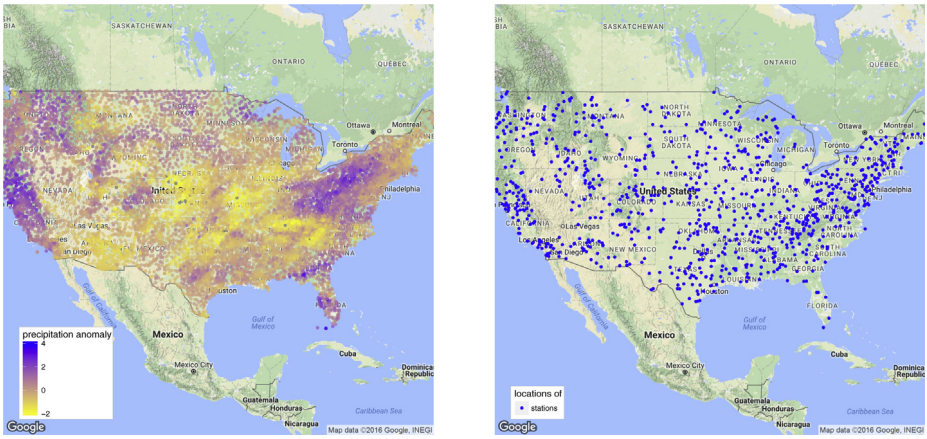

Fig. 12. Location of the 11 000 monitoring stations used to estimate the model parameters (left). Location of the 918 monitoring stations used in the mean square error calculation (right).

standardized monthly total precipitation. As in Liang et al. (2013), we divided the data randomly into two parts. One is composed of 11,000 observations and is used to estimate the model parameters (Fig. 12, left). The other part, with 918 stations, is used to quantify the prediction quality, through the mean square error (Fig. 12, right).

The covariance functions used were the exponential and Matérn, with $B = 25$ and different subsample sizes m. As the generating data model is unknown, the quality of the estimates is compared using the MLE with the complete data as the gold standard. We compare the results obtained using our methods with those obtained given by the RSA estimator of Liang et al. (2013). The estimated parameters are presented in Table 3.

The results show that *RSA*, $\tilde{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\theta}}$ estimate accurately the ratio ϕ/σ^2. Considering the processing time, $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ demand less time than RSA. For hardware composed by 4 cores, the difference is only of few seconds for subsamples $m = 100$ and $m = 300$, and of minutes for $m = 500$ and $m = 700$. When the number of cores increases to 24, the difference also increases. For example, for $m = 700$, the proposed estimators are 8 times faster than RSA and 841 times faster than the MLE.

To calculate the predicted values $\hat{\mathbf{Y}}$ for the 918 stations it is necessary to invert a 11 000 $\times$ 11 000 matrix, requiring a large memory and long processing time. Liang et al. (2013) consider only the observations within a given radius around the points to be predicted, so as to decrease the matrix size to run the kriging algorithm. Similarly, we fix the number of neighbors around the point to be predicted to calculate the covariance matrix of the BLUP estimator. For each of the 918 stations, we ran

Table 3
Comparison of the estimates of MLE, RSA, $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ for the exponential correlation function.

Method	m	β_0	ϕ	σ^2	ϕ/σ^2	τ^2	4 cores (min)	24 cores (min)
MLE	–	0.256	9.999	2.966	3.371	0.049	5634.772	–
$\hat{\boldsymbol{\theta}}$	100	0.089	7.130	2.411	2.957	0.045	0.168	0.069
	300	0.084	9.968	3.239	3.077	0.041	2.648	0.638
	500	0.343	12.046	3.652	3.298	0.053	9.515	2.057
	700	0.334	17.962	5.585	3.209	0.050	45.824	6.683
$\tilde{\boldsymbol{\theta}}$	100	0.192	7.672	2.700	2.841	0.044	–	–
	300	0.063	8.931	3.048	2.93	0.040	–	–
	500	0.265	11.528	3.483	3.309	0.053	–	–
	700	0.214	17.415	5.333	3.216	0.050	–	–
RSA	100	0.149	3.290	0.876	3.756	0.081	0.164	–
	300	0.159	3.160	0.803	3.935	0.058	2.937	–
	500	0.147	2.901	0.821	3.533	0.057	15.473	–
	700	0.150	2.848	0.828	3.440	0.055	53.331	–

Table 4
Comparison of the prediction MSE of the MLE, RSA, $\tilde{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\theta}}$ estimators for the exponential correlation function.

Method	m	Neighbors			
		25	50	100	150
MLE	–	0.0979	0.0976	0.0977	0.0978
$\hat{\boldsymbol{\theta}}$	100	0.0976	0.0972	0.0974	0.0974
	300	0.0989	0.0985	0.0988	0.0988
	500	0.098	0.0977	0.0979	0.0979
	700	0.098	0.0977	0.0979	0.0979
$\tilde{\boldsymbol{\theta}}$	100	0.0975	0.0972	0.0973	0.0973
	300	0.0975	0.0971	0.0973	0.0973
	500	0.0980	0.0977	0.0979	0.0979
	700	0.0980	0.0976	0.0978	0.0979
RSA	100	0.1002	0.0998	0.1002	0.1002
	300	0.0989	0.0985	0.0988	0.0988
	500	0.0984	0.0981	0.0983	0.0983
	700	0.0982	0.0979	0.0981	0.0981

the kriging algorithm considering 25, 50, 100 and 150 nearest neighbors among the 11 000 sampling stations.

Table 4 shows that the number of neighbors does not affect the $\hat{\mathbf{Y}}$ prediction quality. However, an uncommon behavior can be observed, also observed by Liang et al. (2013): the lowest MSE values were obtained when the predictions were based only on the data at the 50 nearest observed locations rather than at 150 neighbors. According to Liang et al. (2013), the cause may be the mispecification of the covariance function or non-stationarity of the data.

Although the MSE values are similar for all estimators we can point out that $\tilde{\boldsymbol{\theta}}$ had the smallest MSE values, followed by $\hat{\boldsymbol{\theta}}$, MLE and RSA. In this example, we can conclude that the spatial subsemble estimators are efficient predictors.

Fig. 13 shows the prediction surfaces of the MLE, $\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}$ and RSA. For the last three estimators, the subsample size is $m = 700$. The prediction surface was evaluated on a regular grid with 10 430 observations. The number of neighbors used to generate the covariance matrix was 25. Comparing the images, there are no substantial differences among the estimates. We can conclude that the predictions based on subsamples are very similar to the surface obtained from the MLE estimator.

To demonstrate the spatial subsemble estimator flexibility, the Matérn covariance function

$$\rho(\|\mathbf{s}_i - \mathbf{s}_j\|; \kappa, \phi) = 2^{\kappa-1}\Gamma(\kappa)^{-1}\left(\frac{\|s_i - s_j\|}{\phi}\right)^{\kappa} K_{\kappa}\left(\frac{\|s_i - s_j\|}{\phi}\right)$$

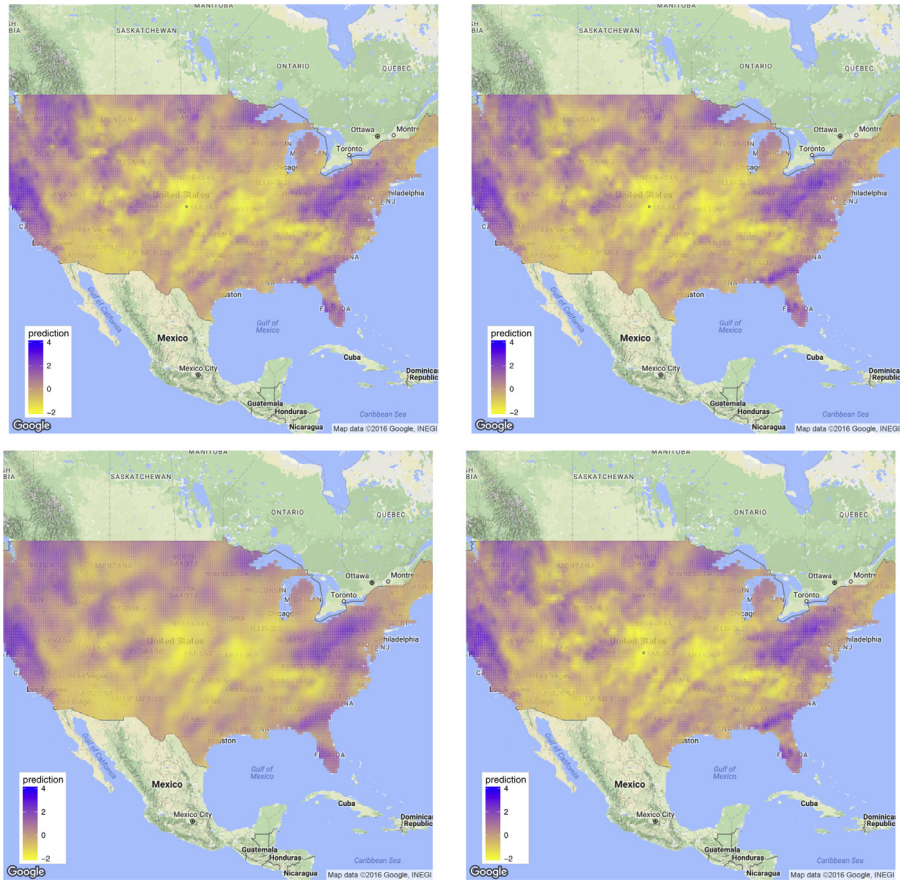

Fig. 13. Prediction surface generated by the MLE (top left), $\hat{\boldsymbol{\theta}}$ (top right), $\tilde{\boldsymbol{\theta}}$ (bottom left) and RSA (bottom right).

Table 5
Comparison of EMV, $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ estimates for the Matérn correlation function, applied to the precipitation data.

Method	m	β_0	ϕ	σ^2	τ^2	κ	4 cores (min)	24 cores (min)
MLE	–	0.121	2.267	0.772	0.050	0.532	21 944.68	
$\hat{\boldsymbol{\theta}}$	300	0.097	2.795	0.721	0.059	1.050	7.208	1.205
	500	0.149	2.350	0.761	0.048	0.629	4.932	21.831
	700	0.088	2.324	0.753	0.045	0.769	21.23	14.270
$\tilde{\boldsymbol{\theta}}$	300	0.090	2.823	0.705	0.054	0.843	–	–
	500	0.148	2.592	0.795	0.048	0.608	–	–
	700	0.080	2.339	0.781	0.047	0.684	–	–

was also used. To generate the results, subsample sizes of 300, 500, and 700 were considered. The number of repetitions is again equal to 25. Table 5 shows that the subsemble estimators have results very close to those of the MLE, specially for the parameters ϕ, σ^2, and τ^2 parameters. For the κ smoothing parameter, the choice of m has a significant influence on the results. It seems necessary to collect a large sample to capture characteristics of the distribution of estimates of this parameter. The estimator $\tilde{\boldsymbol{\theta}}$ gives estimates of κ closer to those from the MLE. With respect to the other parameters, $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ behave similarly.

To reach its estimates, the MLE required 21 944 min, or 15.2 days, of processing time. The estimators $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ were substantially faster: with $m = 100$, it took 1.20 min, while with $m = 700$, it took 14.27 min. Hence, the subsemble estimators can provide huge time saving while maintaining the quality of estimates.

We omit the comparison among the prediction MSE and surfaces calculated using the Matérn correlation function because the results are very similar to Table 4 and they show no visually discernible differences with Fig. 13. Additionally, we can see from Table 5 that the κ estimates of the MLE, $\hat{\boldsymbol{\theta}}^{500}$, $\tilde{\boldsymbol{\theta}}^{700}$, $\tilde{\boldsymbol{\theta}}^{500}$, $\tilde{\boldsymbol{\theta}}^{700}$, are all close to 0.5. When $\kappa = 0.5$, the Matérn function reduces to the exponential function.

5. Consistency of the estimator

There are two approaches to study the asymptotic properties in geostatistics. One approach is the infill asymptotic, when the sampling region is kept fixed but the spatial sampling rate is increased generating ever denser samples. The other approach is the increasing domain asymptotic, when the spatial sampling rate is fixed but the sampling region grows so that more data are collected.

Estimator asymptotic properties are different depending on the approach used. Mardia and Marshall (1984) proved that σ^2 and ϕ from the exponential covariance function can be consistently estimated under the increasing domain asymptotic situation. Under the infill asymptotic context, Stein (2012) showed that neither σ^2 nor ϕ can be estimated consistently although Zhang (2004) showed that the ratio σ^2/ϕ is consistently estimated by the maximum likelihood estimator.

In this section, we only sketch the main results. The proofs follow closely the U-statistics techniques used by Liang et al. (2013) in the case of infill asymptotics, and Politis and Romano (1994) and Politis et al. (1999) in the case of increasing domain asymptotics.

Infill asymptotics

Theorem 5.1 (*Infill Consistency*). *Let $\mathcal{Y} = \{Y(s_1), \ldots, Y(s_n)\}$ be a random sample from a spatial Gaussian field defined on a limited region and represented by (1). Let $\tilde{\boldsymbol{\theta}}$ be the solution of $\binom{n}{m}^{-1}\sum_{i=1}^{\binom{n}{m}} \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{z}_i|\mathbf{s}_i)$. Let Θ be the parametric space of $\boldsymbol{\theta}$ and $\Theta_0 = \{\boldsymbol{\theta}^* \in \Theta : \mathbb{E}[L(\boldsymbol{\theta}^*, \mathbf{Z}|\mathbf{S})] = \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}[L(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{S})]\}$, where $(\mathbf{Z}|\mathbf{S})$ is the random sample of size m from the model (1). Assuming that Θ is compact, then for all $\epsilon > 0$ we have*

$$P(d(\tilde{\boldsymbol{\theta}}, \Theta_0) \geq \epsilon) \to 0, \quad as \quad n \to \infty$$

where $d(\cdot, \cdot)$ is a distance metric.

Proof of 5.1. Because $\mathbf{Z}$ has multivariate normal distribution, $L(\boldsymbol{\theta}, \mathbf{z}|\mathbf{s})$ is continuous and non-negative. Since the normal distribution has finite moments of any order, the assumptions of lemma 2 Liang et al. (2013) are satisfied. Hence, $d(\tilde{\boldsymbol{\theta}}, \Theta_0) \xrightarrow{p} 0$.

Using Theorem 5.1 to calculate $\tilde{\boldsymbol{\theta}}$, we should find a maximum $\binom{n}{m}$ times, making this calculation infeasible. However, according to Lee (1990), in most cases a number of factors in that sum can be omitted without overly inflating the variance estimator; this type of U-statistics is called incomplete. Accordingly, the variance of incomplete U-statistics is always greater than the variance when all observations are used, but its asymptotic efficiency is reasonable even with a B subgroup much smaller than $\binom{n}{m}$. This affirmation is corroborated by Corollary 2.4.1 of Politis et al. (1999).

Increasing domain asymptotics

The proofs and theorems in this section are based on Politis and Romano (1994) and Politis et al. (1999). These authors demonstrated how the subsampling methodology can be extended to the context of time series and random stationary fields. They approximate the sample distribution of a

statistic through statistical values recalculated into small subsets of the data. These recalculated values are adequately normalized and approximate the true sample distribution of the statistics of interest.

Let $\{\mathbf{Y}(\mathbf{s}), \mathbf{s} \in D \subset \mathbb{R}^d\}$ be a random field observed in d dimensions defined on a probability space $(\Omega, \mathcal{A}, P)$. Denote by $\mathbf{s} = (s_1, s_2, \ldots, s_d)$ an arbitrary point in D. For $\mathbf{n} = (n_1, n_2, \ldots, n_d) \in D$, we define the rectangular box

$$E_{\mathbf{n}} = \{\mathbf{s} \in D : 0 < s_k \leq n_k, \text{ for } k = 1, \ldots, d\}.$$

Abusing notation, we let $\mathbf{n} = (n_1, n_2, \ldots, n_d) \in D$ also represent the set of observed sample locations arranged as a grid with $n_1, \ldots, n_d$ distinct values in the coordinate axes and hence $n = \prod_i n_i$. $|E_{\mathbf{n}}|$ denote the Lebesgue measure of the set $E_{\mathbf{n}}$. To represent the increasing domain situation, we will take $\min_i n_i \to \infty$.

Given the $\mathbf{n}$ sample, let $\tilde{\boldsymbol{\theta}}_{\mathbf{n}}$ be the estimator that uses the full data sample: $\tilde{\boldsymbol{\theta}}_{\mathbf{n}} = \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{y})$. Assume that $\tilde{\boldsymbol{\theta}}_{\mathbf{n}}$ and $\boldsymbol{\theta}$ assume values in a separable Banach space, endowed with a norm $\|\cdot\|$. Define $J_{\mathbf{n}, \|\cdot\|}(P)$ as the sample distribution of $\upsilon_{\mathbf{n}} \|\tilde{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}\|$ based on a sample of size $\mathbf{n}$ of (P), where $\upsilon_{\mathbf{n}}$ is a normalization constant. The equation below defines the cumulative distribution function of the statistic considering the complete sample,

$$J_{\mathbf{n}, \|\cdot\|}(y, P) = \mathbb{P}_P\{\upsilon_{\mathbf{n}} \|\tilde{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}\| \leq y\}.$$

Define now the rectangle that represents a subsample. Define the block $Z_{\mathbf{u}} = \{Y(\mathbf{s}), \mathbf{s} \in E_{\mathbf{u},\mathbf{m}}\}$, where

$$E_{\mathbf{u},\mathbf{m}} = \{\mathbf{s} = (s_1, \ldots, s_d) : u_j < s_j \leq u_j + m_j, j = 1, \ldots, d\}\ .$$

Also, let $E_{\mathbf{n}-\mathbf{m}}$ represent the rectangle consisting of the points $\mathbf{s}$ such that $0 < s_k \leq n_k - m_k$ for $k = 1, \ldots, d$. The vector $\mathbf{m}$ indicates the form and size of this rectangle. As $\mathbf{u}$ varies, the block $E_{\mathbf{u},\mathbf{m}}$ is shifted.

Let the calculated value for the MLE statistic with the subsample $Z_{\mathbf{u}}$ be denoted by $\tilde{\boldsymbol{\theta}}_{\mathbf{n},\mathbf{m},\mathbf{u}} = \tilde{\boldsymbol{\theta}}_{\mathbf{m}}(\mathbf{Z}_{\mathbf{u}})$. Following Politis et al. (1999, page 133), the approximation for $J_{\mathbf{n}, \|\cdot\|}(y, \mathbf{P})$ is given by

$$\hat{J}_{\mathbf{n},\mathbf{m},\|\cdot\|}(y) = |E_{\mathbf{n}-\mathbf{m}}|^{-1} \int_{E_{\mathbf{n}-\mathbf{m}}} 1\{\upsilon_{\mathbf{m}}(\|\tilde{\boldsymbol{\theta}}_{\mathbf{n},\mathbf{m},\mathbf{u}} - \tilde{\boldsymbol{\theta}}_{\mathbf{n}}\|) \leq y\} d\mathbf{u}. \tag{11}$$

The theorem supposes that the random field $\{\mathbf{Y}(\mathbf{s})\}$ is homogeneous, meaning that for any set $E \subset D$ and any point $\mathbf{i} \in D$, the joint distribution of the random variables $\{\mathbf{Y}(\mathbf{s}), \mathbf{s} \in E\}$ is identical to the joint distribution of $\{\mathbf{Y}(\mathbf{s}), \mathbf{s} \in E + \mathbf{i}\}$. Furthermore it will be assumed that the random field $\{\mathbf{Y}(\mathbf{s})\}$ satisfies a certain dependence condition: given two points $\mathbf{s} = (s_1, \ldots, s_d)$ and $\mathbf{s}' = (s'_1, \ldots, s'_d)$ in D, define $\delta(\mathbf{s}, \mathbf{s}') = \sup_j |s_j - s'_j|$ and for two sets E_1, E_2 in $\mathbb{R}^d$ denote $\delta(E_1, E_2) = \inf\{\delta(\mathbf{s}, \mathbf{s}') : \mathbf{s} \in E_1, \mathbf{s}' \in E_2\}$. Define the strong mixture coefficients:

$$\begin{aligned} &\alpha_Y(k; l_1, l_2) \equiv \sup_{E_1, E_2 \subset D} \{|P(A_1 \cap A_2) - P(A_1)P(A_2)| : \\ &A_i \in \mathcal{F}(E_i), |E_i| \leq l_i, i = 1, 2, \delta(E_1, E_2) \geq k\}, \end{aligned} \tag{12}$$

where $\mathcal{F}(E_i)$ is a σ-algebra generated by $\{Y(\mathbf{s}), \mathbf{s} \in E_i\}$. We say that the field has weak dependence if $\alpha_Y(k; l_1, l_2)$ converges to zero at some rate when k goes to infinity and l_1, l_2 are fixed or go to infinity.

Other required assumption for subsampling to perform asymptotically is existence of an asymptotic weak limit for $J_{\mathbf{n}, \|\cdot\|}(P)$ as given below:

Assumption 5.2. $J_{\mathbf{n}, \|\cdot\|}(P)$ converges weakly to a limit law $J_{\|\cdot\|}(P)$ with corresponding distribution function $J_{\|\cdot\|}(\cdot, P)$, when $n_i \to \infty$ for $i = 1, \ldots, d$.

With the notation established and the assumptions made, we can present the theorem that states the consistency of the cumulative sample distribution of the statistic based on the subsample $\hat{J}_{\mathbf{n},\mathbf{m},\|\cdot\|}(y)$.

Theorem 5.3 (*Increasing Domain Consistency*). *Suppose that 5.2 holds and that* $\upsilon_{\mathbf{m}}/\upsilon_{\mathbf{n}} \to 0$, $m_i \to \infty$ *e* $n_i \to \infty$, *for* $i = 1, 2, \ldots, d$. *Assume also that there exists a vector* $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_n)$, *depending on* $\mathbf{n}$, *and such that* $2 \leq \Delta_i \leq (n_i - m_i)/m_i$, *for all* $i = 1, \ldots, d$, *as well as* $|\Delta| = \prod_i \Delta_i \to \infty$ *and*

$$|\Delta|\alpha_Y \left(\min_i \left[\frac{n_i - m_i}{\Delta_i} - m_i \right]; (2^{-d}|\Delta| - 1)C(\mathbf{n}, \mathbf{m}, \boldsymbol{\Delta}), 2^d C(\mathbf{n}, \mathbf{m}, \boldsymbol{\Delta}) \right) \to 0 \tag{13}$$

where $C(\mathbf{n}, \mathbf{m}, \boldsymbol{\Delta}) = \prod_i \left(\frac{n_i - m_i}{\Delta_i} + m_i \right)$.

i. *If* y *is a continuity point of* $J_{\|\cdot\|}(\cdot, P)$, *then* $\hat{J}_{\mathbf{n},\mathbf{m},\|\cdot\|}(y) \to J_{\|\cdot\|}(y, P)$ *in probability.*
ii. *If* $J_{\|\cdot\|}(\cdot, P)$ *is continuous, then* $\sup_y |\hat{J}_{\mathbf{n},\mathbf{m},\|\cdot\|}(y) - J_{\mathbf{n},\|\cdot\|}(y, P)| \to 0$ *in probability.*
iii. *Let*

$$c_{\mathbf{n},\mathbf{m},\|\cdot\|}(1-\alpha) = \inf\{y : \hat{J}_{\mathbf{n},\mathbf{m},\|\cdot\|}(y) \geq 1 - \alpha\}.$$

Correspondingly, define

$$c_{\|\cdot\|}(1-\alpha, P) = \inf\{y : J_{\|\cdot\|}(y, P) \geq 1 - \alpha\}.$$

If $J_{\|\cdot\|}(\cdot, P)$ *is continuous at* $c_{\|\cdot\|}(1-\alpha, P)$, *then*

$$Prob_P\{\upsilon_{\mathbf{n}}\|\tilde{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}\| \leq c_{\mathbf{n},\mathbf{m},\|\cdot\|}(1-\alpha)\} \to 1 - \alpha.$$

Then the asymptotic coverage probability under P *of* $\{\boldsymbol{\theta} : \upsilon_{\mathbf{n}}\|\tilde{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}\| \leq c_{\mathbf{n},\mathbf{m},\|\cdot\|}(1-\alpha)\}$ *has nominal level* $1 - \alpha$.

Proof of 5.3. Let $\tilde{\boldsymbol{\theta}}_{\mathbf{n}}$ be the MLE. This estimator is asymptotically normal (Mardia and Marshall, 1984) and so the process has stationary distribution and assumption 5.2 is valid. In addition, m can be chosen such that $m/n = o(1)$ and $m \to \infty$. □

6. Conclusion

In this paper, we proposed two estimators for spatial big data geostatistical analysis. The *spatial subsemble estimators* are examples of the general approach known as *divide and conquer*, in which the estimation problem is divided into three steps: first, small subsets of the data are selected; next, each subset is analyzed separately; finally, the results are aggregated to generate the estimates. The major difference between this new methodology is that each small subset can be analyzed separately, making it possible to use parallel computation. This reduces the time necessary to obtain the estimates, as well as requiring less RAM memory usage. Another advantage is that it is easy to implement.

We carried out a Monte Carlo study to compare the proposed estimators with a methodology indicated to the treatment of massive geostatistical datasets (RSA) and with the golden standard provided by the MLE based on the complete large dataset. The results indicate that the sample distribution of the spatial subsemble and the MLE estimators are similar. It was observed that the quality of the estimators depends on the subsample size, the spatial correlation strength, and the presence of a nugget effect. Larger subsamples produced estimators with smaller bias and variance. The weaker the spatial correlation, the better the estimators' performance. When the nugget effect is not zero, parameter estimation is more difficult, this being shared between the MLE, $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$. In the simulated study, we compared the estimators RSA and subsemble. Our proposed estimator gave better estimates of the covariance function parameters. Of particular importance is the fact that subsemble estimators were faster than RSA, especially when the multiprocessors were used.

Using the Fisher information matrix, we measured the spatial subsemble estimators' variability making it possible to obtain confidence intervals. We also derived the consistency of our estimators under two different asymptotic regimes, the infill and the increasing domain asymptotics. Our estimators can be extended to consider the space–time context where the big data bottleneck becomes more severe.

Acknowledgments

We would like to thank CAPES and CNPq for partial financial support. We also thank the support of the National Supercomputing Center (CESUP) of the Universidade Federal do Rio Grande do Sul (UFRGS) and three anonymous referees for their helpful comments and suggestions.

References

Andrieu, C., Moulines, É., Priouret, P., 2005. Stability of stochastic approximation under verifiable conditions. SIAM J. Control Optim. 44 (1), 283–312.
Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (4), 825–848.
Bühlmann, P., Drineas, P., Kane, M., Lann, M. v. d., 2016. Handbook of Big Data. Chapman and Hall/CRC, New York.
Castrillón-Candás, J.E., Genton, M.G., Yokota, R., 2016. Multi-level restricted maximum likelihood covariance estimation and kriging for large non-gridded spatial datasets. Spat. Stat. 18, Part A, 105–124.
Chen, X., Xie, M., 2014. A split-and-conquer approach for analysis of extraordinarily large data. Statist. Sinica 24 (4), 1655–1684.
Cressie, N., 2015. Statistics for spatial data. John Wiley & Sons, New York.
Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E., 2016. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Amer. Statist. Assoc. 111 (514), 800–812.
Diggle, P.J., Ribeiro Jr., P.J., 2007. Model-Based Geostatistics. Springer, New York.
Finley, A.O., Sang, H., Banerjee, S., Gelfand, A.E., 2009. Improving the performance of predictive process modeling for large datasets. Comput. Statist. Data Anal. 53 (8), 2873–2884.
Fuentes, M., 2007. Approximate likelihood for large irregularly spaced spatial data. J. Amer. Statist. Assoc. 102 (477), 321–331.
Furrer, R., Genton, M.G., Nychka, D., 2006. Covariance tapering for interpolation of large spatial datasets. J. Comput. Graph. Statist. 15 (3), 502–523.
Guha, S., Hafen, R., Rounds, J., Xia, J., Li, J., Xi, B., Cleveland, W.S., 2012. Large complex data: divide and recombine (d&r) with rhipe. Stat 1 (1), 53–67.
Katzfuss, M., 2013. Bayesian nonstationary spatial modeling for very large datasets. Environmetrics 24 (3), 189–200.
Katzfuss, M., 2016. A multi-resolution approximation for massive spatial datasets. J. Amer. Statist. Assoc. URL http://dx.doi.org/10.1080/01621459.2015.1123632.
Kaufman, C.G., Schervish, M.J., Nychka, D.W., 2008. Covariance tapering for likelihood-based estimation in large spatial data sets. J. Amer. Statist. Assoc. 103 (484), 1545–1555.
Kleiner, A., Talwalkar, A., Sarkar, P., Jordan, M.I., 2014. A scalable bootstrap for massive data. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (4), 795–816.
Konomi, B.A., Sang, H., Mallick, B.K., 2014. Adaptive Bayesian Nonstationary Modeling for Large Spatial Datasets Using Covariance Approximations. J. Comput. Graph. Statist. 23 (3), 802–829.
Lee, A.J., 1990. U-Statistics: Theory and Practice. In: Statistics: A Series of Textbooks and Monographs, Taylor & Francis, New York.
Liang, F., Cheng, Y., Song, Q., Park, J., Yang, P., 2013. A resampling-based stochastic approximation method for analysis of large geostatistical data. J. Amer. Statist. Assoc. 108 (501), 325–339.
Lindgren, F., Rue, H., Lindstrom, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (4), 423–498.
Mardia, K.V., Marshall, R.J., 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika 71 (1), 135–146.
Politis, D.N., Romano, J.P., 1994. Large sample confidence regions based on subsamples under minimal assumptions. Ann. Statist. 22 (4), 2031–2050.
Politis, D.N., Romano, J.P., Wolf, M., 1999. Subsampling. Springer, New York.
Ribeiro, P.J., Diggle, P.J., 2001. geoR: a package for geostatistical analysis. R-NEWS (ISSN: 1609-3631) 1 (2), 14–18 URL http://CRAN.R-project.org/doc/Rnews/.
Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Statist. 22 (3), 400–407.
Rue, H., Tjelmeland, H., 2002. Fitting Gaussian Markov random fields to Gaussian fields. Scand. J. Statist. 29 (1), 31–49.
Sang, H., Jun, M., Huang, J.Z., 2011. Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. Ann. Appl. Stat. 5 (4), 2519–2548.
Sapp, S., van der Laan, M.J., Canny, J., 2014. Subsemble: an ensemble method for combining subset-specific algorithm fits. J. Appl. Stat. 41 (6), 1247–1259.
Schifano, E.D., Wu, J., Wang, C., Yan, J., Chen, M.-H., 2016. Online Updating of Statistical Inference in the Big Data Setting. Technometrics 58 (3), 393–403.
Stein, M.L., 2008. A modeling approach for large spatial datasets. J. Korean Stat. Soc. 37 (1), 3–10.
Stein, M.L., 2012. Interpolation of spatial data: some theory for kriging. Springer Science & Business Media, New York.
Stein, M.L., Chi, Z., Welty, L.J., 2004. Approximating likelihoods for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 66 (2), 275–296.
Sun, Y., Stein, M.L., 2016. Statistically and computationally efficient estimating equations for large spatial datasets. J. Comput. Graph. Statist. 25 (1), 187–208.
Tanenbaum, A., 2009. Modern operating systems. Pearson Education, Inc., Amsterdam.

Vecchia, A.V., 1988. Estimation and model identification for continuous spatial processes. J. R. Stat. Soc. Ser. B Stat. Methodol. 50 (2), 297–312.
Zhang, H., 2004. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. J. Amer. Statist. Assoc. 99 (465), 250–261.