Approximate Likelihoods for Spatial Processes

Petrutza Caragea* and Richard L. Smith†

February 9, 2006

Abstract

Maximum likelihood and related techniques are generally considered the best method for estimating the parameters of spatial models, but exact computation of the likelihood is slow when the number of data points is large. For Gaussian models with parametrically specified covariance function, we consider three alternatives to the exact maximum likelihood estimates that are easier to compute. Statistical properties of these estimators are evaluated in two ways, (a) comparing the asymptotic variance of the proposed estimator with that of MLE, (b) assessing how well standard errors computed from the observed information approach (treating the approximate likelihood as if it were an exact likelihood) correspond to the true standard deviations of the estimators. The information sandwich approach is extensively used as out principal theoretical tool for answering these questions. We evaluate the estimators theoretically and by simulation, and consider the application of the method to spatial estimation of rainfall trends across the south-central U.S. Among our three alternatives to exact MLE, the "hybrid method" emerges as the one with the best all-round properties.

Key Words. Big blocks estimator. Hybrid estimator. Kriging. Information sandwich formula. Maximum likelihood. Restricted maximum likelihood. Spatial statistics. Small blocks estimator.

*Department of Statistics, Iowa State University, 125 Snedecor Hall, Ames, IA, 50011-1210. Supported in part by a SAMSI graduate studentship (NSF grant DMS-0112069) and by NSF grant DMS-0084375 while a graduate student at the University of North Carolina.

†Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260; rls@email.unc.edu. Supported in part by NSF grant DMS-0084375. Thanks to Dr. Pavel Groisman of National Climatic Data Center for assistance with the data of Section 5.

1 Introduction

As environmental data sets become larger and more complex, there is a need to develop statistical analysis techniques that handle large data sets efficiently. For example, the current U.S. climatological network maintained by the National Climatic Data Center has over 5,000 stations, and even larger data sets are being created by satellite, radar and other remote sensing devices. Statistical techniques developed in the days when most data sets were much smaller than this cannot usually be applied without modification.

Although there are by now many different stochastic models for spatial processes, those based on Gaussian random fields are among the oldest but still also some of the most commonly applied, see e.g. Journel and Huijbregts (1978), Ripley (1981, 1988), Cressie (1993), Stein (1999), Chilès and Delfiner (1999), Smith (2001). Apart from their direct applicability in many contexts, these models are also used as one of the building blocks in hierarchical analyses of spatial data (Diggle *et al.* 1998, Banerjee *et al.* 2004).

In this paper, we restrict ourselves to models of the form

$$Y \sim N[X\eta, \Sigma(\theta)], \tag{1}$$

where Y is an $N \times 1$ vector of observations, X is an $N \times P$ matrix of covariates, η is a $P \times 1$ vector of unknown regression coefficients, and $\Sigma(\theta)$ is the $N \times N$ covariance matrix, expressed in terms of a finite-dimensional parameter vector θ. We use η for the vector of regression coefficients, rather than the more conventional β, to avoid confusion with the block notation introduced in Section 2. If the process is *stationary* and *isotropic*, the covariance σ_{ij} between any two components Y_i and Y_j is a function of the Euclidean distance d_{ij} between the two observation locations, of which two examples are the *exponential* covariance function,

$$\sigma_{ij} = \sigma^2 \exp\left(-\frac{d_{ij}}{\rho}\right) \tag{2}$$

in which $\theta = (\sigma, \rho)$, σ being a scaling constant and ρ the *range* parameter, and the *Matérn* covariance function

$$\sigma_{ij} = \frac{\sigma}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2} d_{ij}}{\rho}\right)^{\nu} \mathcal{K}_\nu \left(\frac{2\nu^{1/2} d_{ij}}{\rho}\right), \tag{3}$$

where $\mathcal{K}_\nu$ is a modified Bessel function, Γ is the gamma function, and the parameter $\theta = (\sigma, \rho, \nu)$ represent the scale, range and *shape* of the spatial covariance structure. Our general approach, however, does not require stationarity; models for nonstationary processes such as those of Sampson

and Guttorp (1992), Higdon *et al.* (1999) or Fuentes and Smith (2001) may be treated the same way provided the covariance matrix Σ is expressible in terms of a finite-dimensional parameter θ.

Among statistical estimation techniques, for parametric models such as (1) it is natural to consider maximum likelihood and related techniques such as restricted maximum likelihood or REML estimation. Older alternatives, such as estimation based on the sample variogram (Cressie 1993, Chilès and Delfiner 1999), are still widely applied by geoscientists, but the sampling properties of these methods are a matter of speculation in large data sets. Maximum likelihood methods have been known for more than twenty years, e.g. Kitanidis (1983), Mardia and Marshall (1984), and despite some issues such as multimodality (Warnes and Ripley 1987, Mardia and Watkins 1989) are widely accepted. In hierarchical models it is natural to apply Bayesian methods (Diggle *et al.* 1998, Banerjee *et al.* 2004) but these also require calculation of the likelihood function. Therefore, calculating the likelihood function for spatial models such as (1) is of critical importance.

For large data sets, the most critical part of the likelihood calculation is to evaluate the determinant and inverse of the covariance matrix. Theoretically, it is possible to evaluate the determinant and inverse of an $N \times N$ matrix in $O(N^{2.81})$ steps (Aho *et al.* 1974), but most practical algorithms such as Cholesky decomposition require $O(N^3)$ steps. This can be prohibitive if N is large. This motivates us to look for approximations to the likelihood function that require fewer than $O(N^3)$ steps to evaluate, but that still have reasonable statistical properties.

One such scheme was proposed by Vecchia (1988), and may be summarized as follows. Suppose the data vector Y consists of N observations, denoted $Y_1, ..., Y_N$. The ordering of observations is arbitrary. Using p to denote a generic (conditional or unconditional) density, the exact joint density may be written

$$p(Y) \quad = \quad p(Y_1) \prod_{i=2}^{N} p(Y_i \mid Y_1, ..., Y_{i-1}). \tag{4}$$

Vecchia's idea was as follows: suppose in (4) the conditional density $p(Y_i \mid Y_1, ..., Y_{i-1})$ is replaced by $p(Y_i \mid S_i)$,where S_i is some subset of the observations $Y_1, ..., Y_{i-1}$. If $|S_i|$ is not too large, it should be possible to calculate $p(Y_i \mid S_i)$ relatively quickly for each i, and hence derive a computationally efficient approximation to $p(Y)$. In Vecchia's proposal, each S_i consisted of a number of near neighbors of Y_i, though the precise choice of S_i was arbitrary.

Recently Stein *et al.* (2004) generalized Vecchia's idea in a number of ways. They developed a variant of the method to approximate the restricted likelihood function in place of the likelihood function itself. They argued that, rather than evaluate conditional densities one observation at

a time, it might be more efficient to do it in blocks, evaluating conditional densities of the form $p(Y_i, Y_{i+1}, ..., Y_{i+k} \mid S_i)$. They showed that it is not necessarily best to choose S_i consisting only of near neighbors of the observation or observations whose conditional density is being evaluated, since in some contexts, observations further away may have more predictive power. Finally, they developed an "information sandwich" approximation for the covariance matrix of the resulting estimator, similar to what we do in Section 3 below. Although the work of Stein *et al.* is a considerable improvement on Vecchia's original idea, it still suffers from some *ad hoc* features, including both the ordering of the observations and the selection of conditioning sets S_i.

Our approach is based on a different idea for approximating the likelihood. Suppose the observation locations are grouped into blocks of roughly the same size. We describe three methods of approximating the likelihood based on blocks:

1. *Big blocks method.* For each block, compute the block mean. The big blocks likelihood is just the joint density of the block means.

2. *Small blocks method.* For each block, compute the joint density of all observations in that block. The small blocks likelihood is the product of joint densities for all the blocks, in effect treating the blocks as if they were mutually independent.

3. *Hybrid method.* Start with the big blocks likelihood. For each block, compute the joint density of observations in that block, conditional on the block mean. The hybrid likelihood is the big blocks likelihood multiplied by the product of the conditional densities for the blocks. In effect, the hybrid likelihood assumes that the deviations from each block mean, conditional on the block mean itself, are independent across blocks.

The term "big blocks" is intended to convey that the method captures the large-scale properties of the process, but that small-scale (within-block) information is ignored. Similarly, the "small blocks" method captures within-block properties but ignores the between-block dependence. The hybrid method is intended to combine the best features of both methods, hence the name. It is a natural conjecture that the hybrid method is the best of the three, but as we shall see, the comparison is not so simple; from the point of view of efficiency of statistical estimation, the small blocks method is often as good as, or even slightly superior to, the hybrid method, and there are even some situations (admittedly rarer) where the big blocks method is the best of the three.

Before discussing the statistical justification of these methods, we comment briefly on the computational savings to be expected from using any of these methods. Suppose the N observations

are divided into B blocks of roughly K observations per block, so that $N \approx KB$. (In specific calculations to be given later in the paper, we shall generally assume $N = BK$ and each block has exactly K observations, but this is not necessary for the practical application of the method.) Recall that the evaluation of the exact joint density of a set of N observations requires $O(N^3)$ steps.

For the big blocks method, it requires $O(BK)$ steps to compute all the block means, and $O(B^2K^2)$ steps to compute all the block covariances. The final evaluation of the likelihood requires $O(B^3)$ steps. Thus, evaluation of the big blocks likelihood requires $O(B^2K^2 + B^3)$ computational steps.

For the small blocks method, each block likelihood requires $O(K^3)$ steps, and this is repeated B times, requiring $O(BK^3)$ steps in total.

In the hybrid method, the calculation of conditional block likelihoods is no more computationally intensive than the calculation of unconditional block likelihoods for the small blocks method, so the total calculation is the sum of those required for the big blocks and small blocks method, $O(B^2K^2 + B^3 + BK^3)$ steps.

If B grows at a rate between $O(N^{1/2})$ and $O(N^{2/3})$, then all three of these approximate likelihoods may be computed in $O(N^2)$ steps. This differs from the Vecchia and Stein *et al.* methods, where the computational time can be made arbitrarily close to $O(N)$ by choosing the conditioning sets small enough, but even a reduction of computation time from $O(N^3)$ to $O(N^2)$ is a significant saving in large data sets, and greatly extends the potential for fitting Gaussian random field models.

The rest of the paper is laid out as follows. Section 2 defines notation and gives further details of the methods. In this section we also describe some extensions of the method such as REML estimation, estimating the regression parameters η, and spatial interpolation (kriging). Section 3 is the main theoretical development of the paper, presenting an application of the "information sandwich" method to assess the statistical properties of the three methods when they are applied to stationary data on a lattice. Section 4 gives numerical results from the theoretical comparisons of Section 3. Section 5 discusses a practical example, and finally, Section 6 gives a summary of the paper and our conclusions.

2 Details of the methods

Ignoring irrelevant constants, the negative log likelihood function for the model (1) is

$$\ell(\eta, \theta) \;=\; \frac{1}{2}\log|\Sigma(\theta)| + \frac{1}{2}(Y - X\eta)^T\Sigma(\theta)^{-1}(Y - X\eta). \tag{5}$$

If θ and hence the covariance matrix $\Sigma(\theta)$ are known, the usual point estimator of η is the generalized least squares estimator, $\hat{\eta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$, with covariance matrix $(X^T\Sigma^{-1}X)^{-1}$. Therefore, it is usual to concentrate on the estimation of θ, which may be done in two ways. First, if we minimize (5) with respect to η, we obtain the negative log profile likelihood of θ, which modulo a constant is

$$\ell_P(\theta) \;=\; \frac{1}{2}\log|\Sigma(\theta)| + \frac{1}{2}G^2(\theta), \tag{6}$$

where $G^2(\theta) = Y^T\{\Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\}Y$ is the generalized residual sum of squares. The value of θ that minimizes $\ell_P(\theta)$ is the maximum likelihood estimator or MLE of θ.

The second approach is to consider not the full joint density of Y but the joint density of a set of orthogonal contrasts to X. This leads to the restricted likelihood or REML procedure. An alternative formulation, which turns out to be equivalent, is to place a uniform prior density on η and integrate the likelihood derived from (5) with respect to η, which leads to the restricted log likelihood function

$$\ell_R(\theta) \;=\; \frac{1}{2}\log|\Sigma(\theta)| + \frac{1}{2}\log|X^T\Sigma(\theta)^{-1}X| + \frac{1}{2}G^2(\theta), \tag{7}$$

where G^2 is the same as in (6). The REML estimator is then the value of θ that minimizes (7) (Cressie 1993, Smith 2001, Stein *et al.* 2004).

For the moment we focus just on (5), assuming η is known (in which case, without loss of generality, we assume $\eta = 0$). The extensions to (6) and (7) are considered in Section 2.1.

We denote individual observations by Roman indices such as Y_i, Y_j, and blocks by Greek indices, such as b_α to denote the αth block. In general we do not assume all blocks are the same size, and let K_α denote the number of observations in block b_α. Let $\bar{Y}_\alpha$ denote the αth block mean, i.e.

$$\bar{Y}_\alpha = \frac{1}{K_\alpha}\sum_{i\in b_\alpha} Y_i.$$

For blocks $b_\alpha,\ b_\beta$, we have

$$Cov(\bar{Y}_\alpha, \bar{Y}_\beta) \;=\; \frac{1}{K_\alpha K_\beta}\sum_{i\in b_\alpha}\sum_{j\in b_\beta}\sigma_{ij}. \tag{8}$$

Let Σ_{bb} denote the $B \times B$ matrix whose (α, β) entry is given by (8), and let $a_{\alpha\beta}$ denote the (α, β) entry of Σ_{bb}^{-1}. Then the approximate negative log likelihood that is minimized by the big blocks estimator (assuming $\eta = 0$) is

$$\ell_{bb}(\theta) = \frac{1}{2}\log|\Sigma_{bb}(\theta)| + \frac{1}{2}\sum_{\alpha,\beta}\bar{Y}_\alpha\bar{Y}_\beta a_{\alpha\beta} \tag{9}$$

For the small blocks estimator, let $\Sigma_{sb(\alpha)}$ denote the covariance matrix of $\{Y_i,\ i \in b_\alpha\}$. For indices i and j within the same block b_α, let c_{ij} denote the entry of $\Sigma_{sb(\alpha)}^{-1}$ that corresponds to position (i, j). Thus the approximate negative log likelihood to be minimized in this case is

$$\ell_{sb}(\theta) = \frac{1}{2}\sum_{\alpha=1}^{B}\log|\Sigma_{sb(\alpha)}| + \frac{1}{2}\sum_{\alpha=1}^{B}\sum_{i\in b_\alpha, j\in b_\alpha} Y_i Y_j c_{ij}. \tag{10}$$

For the hybrid estimator, we need to compute the conditional joint density of each block given the block mean. To avoid degeneracies, we write each $b_\alpha = b'_\alpha \cup b''_\alpha$ where $|b''_\alpha| = 1$, and let $\Sigma'_{sb(\alpha)}$ denote the unconditional variance of $Y_i,\ i \in b'_\alpha$, given $\bar{Y}_\alpha$. In other words, $\Sigma'_{sb(\alpha)}$ is the same as $\Sigma_{sb(\alpha)}$ with one row and column removed. Then the joint distribution of $\{Y_i,\ i \in b'_\alpha\}$ and $\bar{Y}_\alpha$ is multivariate normal with mean 0 and covariance matrix

$$\begin{pmatrix} \Sigma'_{sb(\alpha)} & \tau_\alpha \\ \tau_\alpha^T & \sigma^2_{bb(\alpha)} \end{pmatrix}$$

where $\sigma^2_{bb(\alpha)} = Var\{\bar{Y}_\alpha\}$ and τ_α is the vector with entries $\tau_i = Cov(Y_i, \bar{Y}_\alpha),\ i \in b'_\alpha$. From standard formulae for conditional multivariate normal distributions, the conditional distribution of $\{Y_i,\ i \in b'_\alpha\}$ given $\bar{Y}_\alpha$ is normal with mean $\tau_\alpha\bar{Y}_\alpha/\sigma^2_{bb(\alpha)}$ and covariance matrix

$$\Sigma_{hyb(\alpha)} = \Sigma'_{sb(\alpha)} - \frac{\tau_\alpha\tau_\alpha^T}{\sigma^2_{bb(\alpha)}}.$$

Let us write $\tau_\alpha/\sigma^2_{bb(\alpha)} = \phi_{(\alpha)}(\theta)$, $\Sigma_{hyb(\alpha)} = \Sigma_{hyb(\alpha)}(\theta)$ to emphasize the dependence on the parameter vector θ. Let $\Sigma^{-1}_{hyb(\alpha)}$ have entries $d_{ij}(\theta)$, indexed by the coefficients $i, j \in b'_\alpha$. Similarly, $\phi_{(\alpha)}$ has entries $\phi_i(\theta),\ i \in b'_\alpha$. Then the estimating function to be minimized by the hybrid estimator is

$$\begin{aligned}\ell_{hyb}(\theta) = {} & \frac{1}{2}\log|\Sigma_{bb}| + \frac{1}{2}\sum_{\alpha}\log|\Sigma_{hyb(\alpha)}| + \frac{1}{2}\sum_{\alpha,\beta}\bar{Y}_\alpha\bar{Y}_\beta a_{\alpha\beta} \\ & + \frac{1}{2}\sum_{\alpha}\sum_{i,j\in b'_\alpha}(Y_i - \phi_i\bar{Y}_\alpha)(Y_j - \phi_j\bar{Y}_\alpha)d_{ij}.\end{aligned} \tag{11}$$

2.1 Regression coefficients, REML estimation, and kriging

The discussion so far has focussed solely on the approximation of the simplest form of log likelihood, when the regression parameter η is either absent or known. However the same difficulties, associated with manipulating a large covariance matrix, also affect the estimation of η, the correction to the log likelihood associated with REML estimation (7), and spatial prediction and interpolation (kriging). In particular, all three require calculation of Σ^{-1}.

Because each of the three approximate likelihoods (9), (10) and (11) is, apart from a constant, the negative log density of a multivariate normal distribution, each may be expressed in the form

$$\ell(\theta) \;=\; -\frac{1}{2}\log|R| + \frac{1}{2}Y^T RY \tag{12}$$

where the matrix R has entries $r_{ij},\ 1 \le i \le N, 1 \le j \le N$. For example, in the case of (11) we have $|R| = -\log|\Sigma_{bb}| - \sum_\alpha \log|\Sigma_{hyb(\alpha)}|$, and the individual entries are given by

$$r_{ij} \;=\; \begin{cases} \frac{a_{\alpha\beta}}{K_\alpha K_\beta} & \text{if } i \in b_\alpha,\ j \in b_\beta,\ \alpha \neq \beta, \\ \frac{a_{\alpha\alpha}}{K_\alpha^2} + \bar{d}_{..} & \text{if } i = j \in b''_\alpha, \\ \frac{a_{\alpha\alpha}}{K_\alpha^2} + \bar{d}_{..} - \bar{d}_{i\cdot} & \text{if } i \in b'_\alpha,\ j \in b''_\alpha, \\ \frac{a_{\alpha\alpha}}{K_\alpha^2} + \bar{d}_{..} - \bar{d}_{i\cdot} - \bar{d}_{j\cdot} + d_{ij} & \text{if } i,j \in b'_\alpha, \end{cases} \tag{13}$$

where

$$\begin{aligned} d_{i\cdot} &= \frac{1}{K_\alpha}\sum_{j \in b'_\alpha} d_{ij}\phi_j, \\ d_{..} &= \frac{1}{K_\alpha^2}\sum_{i,j \in b'_\alpha} d_{ij}\phi_i\phi_j. \end{aligned}$$

Comparison of (11) with (5) shows that the role of R in our approximate likelihood calculations may be equated with that of Σ^{-1} in the exact likelihood. This suggests that R may be taken more generally as an approximation to Σ^{-1}. In particular:

1. *Estimation of regression coefficients.* The exact GLS estimator is $\hat{\eta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$, with covariance matrix $(X^T\Sigma^{-1}X)^{-1}$. We propose replacing Σ^{-1} by R to get an approximate GLS estimator.

2. *REML estimation.* Formula (6) requires that we estimate $\log|\Sigma|$, $\log|X^T\Sigma^{-1}X|$ and G^2. We approximate $\log|\Sigma|$ by $-\log|R|$, as previously noted. Both $\log|X^T\Sigma^{-1}X|$ and G^2 depend on Σ on through Σ^{-1}; in each case, we propose substituting R for Σ^{-1} to obtain an approximate value.

3. *Kriging.* Suppose we are interested in predicting or interpolating a value of the random field Y_0 with mean $x_0^T\eta$ and variance σ_0^2. Suppose τ is the vector of cross-correlations between Y_0 and Y. We assume x_0, σ_0^2 and τ are known, though each may be a function of the unknown parameter vector θ. According to the theory of universal kriging (Ripley 1981, Cressie 1993, Chilès and Delfiner 1999, Stein 1999, Smith 2001), the best linear predictor of Y_0 is $\hat{Y}_0 = \lambda^T Y$, where $\lambda = \Sigma^{-1}\tau + \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau)$, and the associated mean squared prediction error is $\sigma_0^2 - \tau^T\Sigma^{-1}\tau + (x_0 - X^T\Sigma^{-1}\tau)^T(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau)$. We propose evaluating both of these formulae approximately, by substituting R for Σ^{-1}.

3 Application of the "Information Sandwich" to assess statistical properties of the estimators

This section and the next focus on theoretical properties of the methods. For certain special cases of spatial processes defined on a one- or two-dimensional lattice, we have been able to calculate rigorous asymptotic efficiencies (Caragea and Smith 2005). Here, we adopt a more intuitive approach based on the "information sandwich" formula for asymptotic variances of estimators defined by estimating equations. We concentrate on the case where the regression parameter η is known (without loss of generality, $\eta = 0$) and focus on the estimation of the spatial covariance parameters θ.

For stationary processes on a lattice, there are already known ways of approximating the likelihood efficiently (e.g. Whittle (1954), Guyon (1982)) or of simplifying the exact likelihood computations (Zimmerman 1989). The reason we are confining our theoretical calculations to lattices is that we believe stationary lattice processes form a sufficiently rich test-bed of examples that we can use to evaluate the properties of our estimators, combined with the fact that it is possible greatly to speed up the Information Sandwich calculations in the case of a stationary lattice process. For actual applications, we do not envision that they be confined to processes on a lattice (see Section 5 for a practical example).

The theoretical discussion focusses on two questions:

Problem 1. What are the efficiencies of our approximate maximum likelihood procedures, as assessed by the ratios of the variances of our parameter estimates to those ofr maximum likelihood? (This is also known as relative efficiency, but in this paper, we omit the word "relative".)

Problem 2. Suppose we estimated the variances of the parameter estimates by the "direct method" of inverting the observed information matrix, ignoring the fact that we are not using the true likelihood function. What are the ratios of the estimation variances calculated by this method to the true estimation variances?

Both questions may be answered (approximately) through an application of the "information sandwich" (IS) method to characterize the sampling variability of our three approximate maximum likelihood estimators. The general principles of the IS approach have been discussed by numerous authors such as Heyde (1997), and were also extensively used by Stein *et al.* (2004).

Our viewpoint of Problem 2 above is different from that of Stein *et al.* (2004). They argued against estimating standard errors simply by inverting the approximate observed information matrix; instead, they proposed a variant on the IS formula for that problem, using a sampling technique to evaluate a subset of the terms in cases where exact evaluation of the IS formula was too cumbersome. Our viewpoint is that most spatial statisticians would not want to go to that much trouble; if we are viewing our estimators as approximate maximum likelihood estimators, then it is natural to want to reap the other benefits of a maximum likelihood approach, including automatic calculation of the standard errors. We view our results on Problem 2 as providing some justification for this viewpoint, while at the same time, in cases where the information matrix approach substantially underestimates the true variances, providing some caution about the limitations of this approach.

3.1 Outline of the IS approach

We assume a stationary Gaussian process $\{Y_i,\ i = 1, ..., N\}$, where i is a two-dimensional index, written (i_1, i_2) when we want to distinguish the two components, with $1 \le i_1 \le N_1,\ 1 \le i_2 \le N_2$, where $N = N_1 N_2$. Suppose $N = BK$ where $B = B_1 B_2$ is the number of blocks, arranged as a $B_1 \times B_2$ array, and suppose each block consists of $K = K_1 K_2$ observations, arranged as a $K_1 \times K_2$ array. We assume $EY_i = 0$ and write $\sigma_{ij}(\theta) = E\{Y_i Y_j\}$, assumed to depend only on $i - j$ by stationarity. We write $\theta_r,\ \theta_s$, etc., for the individual components of θ.

In applying the information sandwich formula to an estimator defined by minimizing the estimating function $S = S(Y_1, ..., Y_N; \theta)$, we write $H(\theta)$ for the matrix with (r, s) entry

$$E\left\{\frac{\partial S}{\partial \theta_r}\frac{\partial S}{\partial \theta_s}\right\}$$

and $W(\theta)$ for the matrix with (r, s) entry

$$E\left\{\frac{\partial^2 S}{\partial\theta_r \partial\theta_s}\right\}.$$

The approximate covariance matrix of the estimator $\tilde{\theta}$, that minimizes S, is then $W(\theta)^{-1}H(\theta)W(\theta)^{-1}$.

Before proceeding to a detailed discussion of individual estimators, we note three general points:

1. $Cov\{Y_iY_j, Y_kY_\ell\} = \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}$. This is used in computing variances of quadratic forms in the Y variables.

2. As is well known from elementary likelihood theory, under mild regularity conditions we have

$$E\left\{\frac{\partial}{\partial\theta_r}\log f(Y;\theta)\right\}_{\theta=\theta_0} = 0$$

whenever f is the density of a random vector Y with true parameter θ_0. This property may also be expressed by saying that the set of first-order partial derivatives of the log likelihood form an unbiased set of estimating equations for θ. Each of our proposed estimating functions $S = \ell_{bb},\ \ell_{sb}$ and ℓ_{hyb} is a sum of negative log conditional or unconditional likelihood functions for different portions of the data, and therefore has the same unbiasedness property, $E\{\partial S/\partial\theta_r\} = 0$ at the true value $\theta = \theta_0$.

3. If $Y \sim N[\mu(\theta), \Sigma(\theta)]$ and the parameters θ are estimated by maximum likelihood, then the Fisher information matrix has (r, s) entry

$$\frac{\partial\mu^T}{\partial\theta_r}\Sigma^{-1}\frac{\partial\mu}{\partial\theta_s} + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_r}\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_s}\right), \tag{14}$$

see, e.g. Smith (2001), Section 6.5.3. Noting again that each S can be written as a sum of negative log likelihood for different portions of the data, it follows at once that the entries of $W(\theta)$ can be expressed as a sum of terms of the form (14).

Once these formulae have been evaluated we apply them as follows:

Problem 1. The efficiency of the estimators compared with maximum likelihood is computed by comparing the approximate covariance matrix $W^{-1}HW^{-1}$ with the inverse of the true Fisher information matrix, say W_0^{-1}. In particular, the ratios of the diagonal entries of the two matrices are the approximate efficiencies of the individual parameter estimators compared with exact MLE.

Problem 2. The ratios of the diagonal entries of $W^{-1}HW^{-1}$ are compared with those of W^{-1} to determine to what extent the variances would be incorrectly estimated if we used the direct method (inverting the approximate observed information matrix) as opposed to the IS method. In this context, a ratio greater than 1 would indicate that the direct method underestimates the true sampling variance of the approximate MLE. We don't do this for the big blocks estimator, because in that case there is no need to distinguish the IS and direct methods: the big blocks estimator uses the exact likelihood based on the block means, and the IS method coincides with W^{-1} in this case.

3.2 Big blocks estimator

As in Section 2, we denote by Σ_{bb} the covariance matrix of the block means. It follows that the (r, s) entry of the Fisher information matrix is

$$\frac{1}{2}\mathrm{tr}\left(\Sigma_{bb}^{-1}\frac{\partial\Sigma_{bb}}{\partial\theta_r}\Sigma_{bb}^{-1}\frac{\partial\Sigma_{bb}}{\partial\theta_s}\right). \tag{15}$$

The inverse of the matrix with entries (15) is therefore the approximate variance-covariance matrix of the big blocks estimator.

3.3 Small blocks estimator

Let Σ_{sb} denote the covariance matrix of $\{Y_i,\ i \in b_\alpha\}$, which is independent of α by stationarity and regularity of the lattice. Since the log likelihood (10) is a sum of log likelihoods for the individual blocks, it follows from (14) that

$$E\left\{\frac{\partial^2 S}{\partial\theta_r\partial\theta_s}\right\} = \frac{B}{2}\mathrm{tr}\left(\Sigma_{sb}^{-1}\frac{\partial\Sigma_{sb}}{\partial\theta_r}\Sigma_{sb}^{-1}\frac{\partial\Sigma_{sb}}{\partial\theta_s}\right) \tag{16}$$

so we use (16) to define the matrix $W(\theta)$.

To compute the matrix $H(\theta)$, let us write

$$T_{\alpha,r} = \frac{1}{2}\sum_{i\in b_\alpha, j\in b_\alpha} Y_iY_j\frac{\partial c_{ij}}{\partial\theta_r} + \frac{1}{2}\frac{\partial}{\partial\theta_r}\log|\Sigma_{sb}|$$

which has mean 0 by Property 2 of Section 3.1. By (10), $\frac{\partial\ell_{sb}}{\partial\theta_r} = \sum_\alpha T_{\alpha,r}$. We also note that

$$Cov(T_{\alpha,r}, T_{\beta,s}) = \frac{1}{4}\sum_{i\in b_\alpha, j\in b_\alpha, k\in b_\beta, \ell\in b_\beta}\frac{\partial c_{ij}}{\partial\theta_r}\frac{\partial c_{k\ell}}{\partial\theta_s}(\sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}) \tag{17}$$

which (for given $\alpha,\ \beta,\ r,\ s$) can be computed in $O(K^4)$ steps. We also note that, by stationarity of the process and regularity of the blocks, (17) depends on the block indices α and β only through their vector difference $\alpha - \beta$. This property is critical in defining the next step.

Consider a specific configuration of (α, β) for which the coordinates satisfy $\beta_1 = \alpha_1 + \chi_1, \beta_2 = \alpha_2 + \chi_2$, $1 - B_1 \leq \chi_1 \leq B_1 - 1$, $1 - B_2 \leq \chi_2 \leq B_2 - 1$. Within the $B_1 \times B_2$ array of blocks, this configuration occurs $(B_1 - |\chi_1|)(B_2 - |\chi_2|)$ times. Therefore

$$Cov\left(\sum_{\alpha=1}^{B} T_{\alpha,r}, \sum_{\beta=1}^{B} T_{\beta,s}\right) = \sum_{\alpha=1}^{B}\sum_{\beta=1}^{B} Cov\left(T_{\alpha,r}, T_{\beta,s}\right) \tag{18}$$

$$= \sum_{\chi_1=1-B_1}^{B_1-1}\sum_{\chi_2=1-B_2}^{B_2-1} (B_1 - |\chi_1|)(B_2 - |\chi_2|) Cov\left(T_{0,r}, T_{\chi,s}\right) \tag{19}$$

which, in combination with (17), requires a total of $O(K^4 B)$ calculations for each (r, s) pair.

Formulae (17) and (19) together define the (r, s) entry of the matrix $H(\theta)$. Since (16) defines the matrix $W(\theta)$, the desired asymptotic covariance matrix for the small blocks estimator then follows from the information sandwich formula, $W(\theta)^{-1} H(\theta) W(\theta)^{-1}$.

Note that the computational saving from writing the formula in the form (19), as compared with evaluating (18) directly, is to reduce a sum over B^2 terms to one over $(2B_1 - 1)(2B_2 - 1)$ terms. For example, in the case $B_1 = B_2 = 9$, considered in one of our main examples below, the computational saving is a factor of $(81/17)^2$, or about 23. Given that the formula typically takes several minutes to compute even via (19), this is a significant practical saving.

3.4 Hybrid estimator

The estimating function for the hybrid estimator is given by (11). For the case of a stationary process on the lattice, we note two simplifications: (a) the matrix $\Sigma_{hyb(\alpha)}$ is the same for all blocks and will henceforth be written Σ_{hyb}; (b) the coefficients $\phi_i,\ \phi_j,\ d_{ij}$ depend only on the position of coordinates i and j within the αth block, and not otherwise on α.

To compute the (r, s) entry of $W(\theta)$ in this case, we again apply Property 3 of Section 3.1 (first conditionally on each $\bar{Y}_\alpha$, and then taking the expectation with respect to $\bar{Y}_\alpha$) to get

$$\frac{1}{2}\mathrm{tr}\left(\Sigma_{bb}^{-1}\frac{\partial \Sigma_{bb}}{\partial \theta_r}\Sigma_{bb}^{-1}\frac{\partial \Sigma_{bb}}{\partial \theta_s}\right) + B\sigma_{bb}^2 \frac{\partial \phi^T}{\partial \theta_r}\Sigma_{hyb}^{-1}\frac{\partial \phi}{\partial \theta_s} + \frac{B}{2}\mathrm{tr}\left(\Sigma_{hyb}^{-1}\frac{\partial \Sigma_{hyb}}{\partial \theta_r}\Sigma_{hyb}^{-1}\frac{\partial \Sigma_{hyb}}{\partial \theta_s}\right). \tag{20}$$

The remaining task is to compute covariances of $\partial S/\partial\theta_r$ and $\partial S/\partial\theta_s$ where

$$\frac{\partial S}{\partial \theta_r} = \frac{1}{2}\sum_{\alpha,\beta} \bar{Y}_\alpha \bar{Y}_\beta \frac{\partial a_{\alpha\beta}}{\partial \theta_r} - \sum_{\alpha}\sum_{i,j\in b'_\alpha} \frac{\partial \phi_i}{\partial \theta_r}\bar{Y}_\alpha (Y_j - \phi_j \bar{Y}_\alpha) d_{ij}$$

$$+\frac{1}{2}\sum_{\alpha}\sum_{i,j\in b'_\alpha}(Y_i-\phi_i\bar{Y}_\alpha)(Y_j-\phi_j\bar{Y}_\alpha)\frac{\partial d_{ij}}{\partial\theta_r}+\frac{\partial}{\partial\theta_r}\left\{\frac{1}{2}\log|\Sigma_{bb}|+\frac{B}{2}\log|\Sigma_{hyb}|\right\}. \quad (21)$$

Let

$$\begin{aligned} U_r &= \frac{1}{2}\sum_{\alpha,\beta}\bar{Y}_\alpha\bar{Y}_\beta\frac{\partial a_{\alpha\beta}}{\partial\theta_r}, \\ T^*_{\alpha,r} &= -\sum_{i,j\in b'_\alpha}\frac{\partial\phi_i}{\partial\theta_r}\bar{Y}_\alpha(Y_j-\phi_j\bar{Y}_\alpha)d_{ij}+\frac{1}{2}\sum_{i,j\in b'_\alpha}(Y_i-\phi_i\bar{Y}_\alpha)(Y_j-\phi_j\bar{Y}_\alpha)\frac{\partial d_{ij}}{\partial\theta_r}. \end{aligned}$$

Then the matrix $H(\theta)$ has entries of the form

$$Cov\left(U_r+\sum_\alpha T^*_{\alpha,r}, U_s+\sum_\beta T^*_{\beta,s}\right). \quad (22)$$

The expression (22) is the sum of four terms:

1. $Cov(U_r, U_s)$ is of the form

$$\frac{1}{4}\sum_{\alpha,\beta,\gamma,\delta}\frac{\partial a_{\alpha\beta}}{\partial\theta_r}\frac{\partial a_{\gamma\delta}}{\partial\theta_s}\left\{Cov(\bar{Y}_\alpha,\bar{Y}_\gamma)Cov(\bar{Y}_\beta,\bar{Y}_\delta)+Cov(\bar{Y}_\alpha,\bar{Y}_\delta)Cov(\bar{Y}_\beta,\bar{Y}_\gamma)\right\}. \quad (23)$$

2. $Cov(U_r, T^*_{\beta,s})$ is of the form

$$\frac{1}{2}\sum_{\gamma,\delta}\frac{\partial a_{\gamma\delta}}{\partial\theta_r}Cov\left[\bar{Y}_\gamma\bar{Y}_\delta,\left\{-\sum_{i,j\in b'_\beta}\frac{\partial\phi_i}{\partial\theta_s}\bar{Y}_\beta(Y_j-\phi_j\bar{Y}_\beta)d_{ij}+\frac{1}{2}\sum_{i,j\in b'_\beta}(Y_i-\phi_i\bar{Y}_\beta)(Y_j-\phi_j\bar{Y}_\beta)\frac{\partial d_{ij}}{\partial\theta_s}\right\}\right]. \quad (24)$$

3. $Cov(U_s, T^*_{\alpha,r})$ is computed the same way as (24).

4. Using the same trick as for the small blocks estimator, we write

$$Cov\left(\sum_{\alpha=1}^{B}T^*_{\alpha,r},\sum_{\beta=1}^{B}T^*_{\beta,s}\right)=\sum_{\chi_1=1-B_1}^{B_1-1}\sum_{\chi_2=1-B_2}^{B_2-1}(B_1-|\chi_1|)(B_2-|\chi_2|)Cov\left(T^*_{0,r},T^*_{\chi,s}\right). \quad (25)$$

Finally, we put (23), (24) and (25) together to deduce $H(\theta)$. Since we also have $W(\theta)$ defined from (20), the approximate covariance matrix of the hybrid estimator is then defined by the information sandwich formula.

4 Results

In this section, we apply the calculations of Section 3 to three examples.

Our first example is a one-dimensional time series of AR(1) form, denoted by $Y_i = Y_{i-1} + \theta e_i$ with e_i independent $N(0,1)$. This is of course an artificial example, since for this model the MLE is computable analytically, but it is used by Caragea and Smith (2005) as an example for which the asymptotic efficiency of all three methods is computable analytically. We let $N = 500$ observations divided into $B = 50$ blocks of size $K = 10$. Table 1 shows the asymptotic efficiency of the estimation of θ, as well as the IS/direct ratios, calculated by the method of Section 3. The efficiencies are low for the big blocks method except when θ is near 1, but they are above 90% for both the small blocks and hybrid methods. Also shown are the IS/direct ratios, which are very close to 1 except when $\theta = \pm 0.75$.

θ	Big Blocks Efficiency	Small Blocks Efficiency	Hybrid Efficiency	Small Blocks IS/Direct	Hybrid IS/Direct
–0.750	.00538	.92595	.92267	1.25	1.26
–0.250	.08999	.91329	.91373	1.002	1.003
–0.010	.15983	.90182	.90280	1.000	1.000
0.010	.16684	.90182	.90281	1.000	1.000
0.250	.27301	.91329	.91409	1.002	1.001
0.750	.73896	.92595	.91800	1.246	1.177

Table 1. Efficiency of estimators for AR(1) model with $B = 50,\ K = 10$.

The next example is designed to compare our theoretical results with simulations, for the particular case of a spatial process with exponential covariances on a 20×20 lattice. We assume the distance between neighboring lattice points is 1, and a range of either 0.5 or 1.5. We also assume square blocks with either $B = 100,\ K = 4$ or $B = 25,\ K = 16$.

For this example, we have computed approximate efficiencies for all three estimators, using the IS formula, as well as the IS/Direct ratios for the small blocks and hybrid methods. Corresponding simulation results are in parentheses, computed from 5000 replications. In the case of the IS/Direct ratios, the simulated results represent the mean of the ratios between the true variance of the estimator (estimated from Monte Carlo replications) to the variance that is estimated using the approximate information matrix — this is the real quantity of interest, that the IS/Direct ratio is trying to approximate. We have also computed standard errors for these quantities to take account of Monte Carlo variability in the simulations; these are not shown in the table, but overall they confirm that the differences between theoretical and simulated results can be attributed to

Monte Carlo variability. We have not attempted more extensive simulation studies because of the computational expense that they involve.

In Table 2, the IS/Direct ratios are generally between 1 and 2, though it is notable that in cases where the ratios are substantially greater than 1 (those with true range 1.5), the IS/Direct ratio is closer to 1 for the hybrid method than for the small blocks method, which may be a reason to prefer the hybrid method. As far as efficiency is concerned, there is not much to choose between the small blocks and hybrid methods.

True range and scale	B, K	Method	Efficiency for range	Efficiency for scale	IS/Direct for range	IS/Direct for scale
0.5,1	100, 4	BB	.172 (.124)	.118 (.134)	N/A	N/A
0.5,1	100, 4	SB	.572 (.503)	1.000 (.999)	1.02 (0.89)	1.04 (1.06)
0.5,1	100, 4	HYB	.665 (.611)	1.000 (.998)	.997 (0.89)	1.03 (1.04)
1.5,1	100, 4	BB	.467 (.426)	.778 (.754)	N/A	N/A
1.5,1	100, 4	SB	.779 (.776)	.949 (.966)	1.63 (1.50)	2.10 (1.92)
1.5,1	100, 4	HYB	.813 (.784)	.964 (.955)	.98 (.91)	1.15 (1.07)
0.5,1	25, 16	BB	.011 (0.015)	.003 (.011)	N/A	N/A
0.5,1	25, 16	SB	.818 (.797)	1.000 (.999)	1.01 (.95)	1.02 (1.03)
0.5,1	25, 16	HYB	.823 (.800)	1.000 (.998)	1.01 (.95)	1.02 (1.03)
1.5,1	25, 16	BB	.090 (.061)	.085 (.056)	N/A	N/A
1.5,1	25, 16	SB	.886 (.887)	.937 (.954)	1.39 (1.33)	1.52 (1.41)
1.5,1	25, 16	HYB	.880 (.842)	.935 (.925)	1.23 (1.22)	1.23 (1.29)

Table 2. Efficiency of estimators for exponential model with unknown range and scale parameters; 20×20 lattice.

Now we consider a larger spatial lattice, based on a 27×27 lattice divided into 81 3×3 blocks. We consider two models, the exponential and Matérn models of (2) and (3).

Tables 3 and 4 present the results for both models where the range parameter is one of 3, 9 or 27 (assuming the distance between neighboring lattice points is one unit), and in the Matérn case, the Matérn scale parameter ν is either 1 or 0.1, the latter representing a case where there is a near-discontinuity in the variogram at the origin.

Here are a few qualitative conclusions we can draw from these tables:

1. The big blocks estimator is again inefficient for when the true range is small, but improves dramatically when the range is large; for range 27 (so that the range of spatial covariance is comparable with that of the data set) the big blocks estimator is sometimes the best of the three, though not for estimating ν in the Matérn covariance (which is logical, since ν is a parameter which reflects the small-scale variability of the process).

2. In most cases, the small blocks and hybrid estimators are comparable in efficiency, with no clear preference for one over the other. The one exception to this is the Matérn model with $\nu = 0.1$; in this case, the hybrid estimator appears clearly superior to the small blocks estimator in most cases.

3. The IS/Direct ratios are sometimes much bigger than 1, implying that variance estimators based on inverting the approximate information matrix will seriously underestimate the true variances. However, in every instance where this happens, the IS/Direct ratio is much smaller for the hybrid estimator than for the small blocks estimator. It is still debatable whether a ratio of 2 of 3 is acceptable in a practical estimator, but the message seems to be that if we propose to use the direct method to approximate the variance of an estimator, we get a more accurate approximation in the case of the hybrid estimator.

True range and scale	Method	Efficiency for range	Efficiency for scale	IS/Direct for range	IS/Direct for scale
3,1	BB	.44704	.87855	N/A	N/A
3,1	SB	.83511	.87175	2.91	3.44
3,1	HYB	.81079	.85079	1.45	1.64
9,1	BB	.75813	.93191	N/A	N/A
9,1	SB	.72408	.73858	10.99	12.70
9,1	HYB	.76690	.77747	1.88	1.98
27,1	BB	.90026	.95448	N/A	N/A
27,1	SB	.71722	.71900	32.06	36.38
27,1	HYB	.77195	.77435	1.99	2.03

Table 3. Efficiency of estimators for exponential model with unknown range and scale parameters; 27×27 lattice divided into 3×3 blocks.

True range, scale and shape	Method	Efficiency for range	Efficiency for scale	Efficiency for shape	IS/Direct for range	IS/Direct for scale	IS/Direct for shape
3,1,1	BB	.38638	.22566	.00552	N/A	N/A	N/A
3,1,1	SB	.67215	.87884	.47059	1.67	2.64	1.30
3,1,1	HYB	.61722	.81863	.47753	1.47	1.80	1.32
9,1,1	BB	.38325	.84483	.02399	N/A	N/A	N/A
9,1,1	SB	.58047	.71886	.42485	4.91	9.79	1.78
9,1,1	HYB	.62819	.74415	.48413	2.22	2.69	1.75
27,1,1	BB	.53529	.88697	.05428	N/A	N/A	N/A
27,1,1	SB	.45077	.59332	.29222	16.28	34.96	3.72
27,1,1	HYB	.70594	.77619	.44586	2.43	3.07	2.37
3,1,.1	BB	.80723	.06917	.07704	N/A	N/A	N/A
3,1,.1	SB	.53118	.96192	.34166	1.37	1.96	1.02
3,1,.1	HYB	.90968	.97333	.77054	1.06	1.09	1.08
9,1,.1	BB	.85495	.41777	.10820	N/A	N/A	N/A
9,1,.1	SB	.62836	.86186	.32750	2.31	6.39	1.03
9,1,.1	HYB	.90449	.92904	.80421	1.14	1.14	1.11
27,1,.1	BB	.89971	.83043	.12496	N/A	N/A	N/A
27,1,.1	SB	.67707	.82933	.31381	3.63	16.83	1.04
27,1,.1	HYB	.89021	.89893	.81865	1.24	1.22	1.11

Table 4. Efficiency of estimators for Matérn model with unknown range, scale and shape parameters; 27×27 lattice divided into 3×3 blocks.

5 An example

In this section we discuss a practical example, designed to illustrate how well the methods perform in a real-data context.

In recent years, climatologists have taken great interest in possible trends in rainfall frequencies and amounts (Karl and Knight 1998). Next to the steady rise in temperatures known as global warming, rainfall trends form one of the best indicators of global climate change. However (as with temperature changes) the actual measured trend varies considerably at different locations. The present example is part of a larger investigation involving spatial interpolation of precipitation trends.

For this study, we have taken 540 rainfall stations across several states in the south-central US; the stations are shown in Fig. 1, where for later comparison we have subdivided them into four subsets corresponding to different subregions. The stations are taken from a larger dataset complied by the National Climatic Data Center; for this example, the number of stations was chosen to be near the upper limit of what is easily computable by exact maximum likelihood.

At each station, monthly totals were taken from 1965 to 1985. A linear trend was estimated by simple linear regression using the model $y_t = at + b_{m(t)} + e_t$ where y_t is observed rainfall total in month t, a is the coefficient of linear trend, $m(t)$ is the month during which t falls (Month 1 is January, Month 2 is February, etc.), $b_1, ..., b_{12}$ are 12 individual month effects and e_t is random error. Only the estimated trends $\hat{a}$ for each station are used in the subsequent spatial analysis; a plot of these is shown in Fig. 2. The plot suggests some spatial coherence in the trends, but does not suggest any systematic spatial pattern that might be modeled by including an $X\eta$ component in (1); in subsequent analysis, we consider only models in which $X\eta$ is reduced to a constant unknown mean, but we explore different possibilities for the spatial dependence matrix $\Sigma(\theta)$.

Some initial investigations were undertaken into the correct form of spatial model. Latitudes and longitudes (measured in degrees) were used to define the spatial coordinates of each station. Plots of the sample variogram (see, e.g., Cressie (1993) for discussion of variograms) suggested the typical pattern found in geostatistics, with a variogram increasing at small distances but levelling off to a "sill", that could be fitted using the exponential or Matérn models, (2) or (3). Directional variograms are often used as a diagnostic for anisotropy; in this case, there was no evidence that the variogram increased more rapidly in one direction than another. Comparison of different parametric models suggested that the exponential model (2) fitted as well as any, but that it was necessary to

include a "nugget effect". Hence the actual model fitted was

$$\sigma_{ij} = \begin{cases} e^{\alpha} & \text{if } d_{ij} = 0, \\ (1-\phi)e^{\alpha - d_{ij}/\rho} & \text{if } d_{ij} > 0, \end{cases} \tag{26}$$

where α is the log-sill (log transformation taken to improve numerical stability), ρ is the range and $\phi \in [0, 1]$ the nugget:sill ratio.

There remains some doubt, however, about whether a single model of the form (26) fits the data throughout the region of interest. To investigate this issue, separate models were also fitted to each of the four subregions indicated in Fig. 1. The initial discussion is based on a single model fitted to all 540 stations, but subsequently we consider the effect of dividing the region into four subregions.

Method	α	ρ	ϕ
MLE	–4.32091 (0.20199)	0.82742 (0.24412)	0.40134 (0.06670)
SB	–4.51278 (0.17630)	1.28372 (0.45045)	0.48709 (0.06527)
HYB	–4.40432 (0.20204)	0.95808 (0.30594)	0.43953 (0.06726)

Table 5. Estimates of the three parameters in model (26), fitted to the full dataset, by exact maximum likelihood (MLE) and the small blocks (SB) and hybrid (HYB) methods; the latter were based on a subdivision into 60 blocks of size 9. Standard errors are in parentheses. In the case of the SB and HYB methods, standard errors were computed by the direct method, i.e. inversion of the approximate observed information matrix.

Table 5 shows the exact maximum likelihood estimator (MLE) for model (26), together with the small blocks (SB) and hybrid (HYB) estimators, when the region is divided into 60 blocks of size 9. Also shown are standard errors (for SB and HYB, computed by the "direct" method). We did not consider the big blocks method because in this example, with the estimated range parameter ρ much smaller than the diameter of the region, it seemed obvious that this would not be competitive. Table 5 suggests that both the SB and HYB estimates are close to the MLE, though in this example, HYB is closer to MLE than SB for all three parameters. The standard error estimates are also comparable across all three estimation methods.

One can argue that the real purpose of such a study is not how well we can estimate the spatial model parameters, but how well it is possible to interpolate the random field using those parameters. To investigate this, Fig. 3 shows the result of spatial interpolation (kriging) using all three estimators, including the nugget effect. For ease of plotting, the interpolation is performed

on a 30×30 grid. The three interpolated surfaces, appear, visually, almost identical.

To investigate further the effect of the different interpolation schemes, a cross-validation procedure was carried out. Each station in turn was deleted and its value predicted from the remaining stations. A cross-validated mean squared prediction error was computed. This was repeated for each of the three estimation methods. As a straw-man comparison, we also performed the same exercise where the "interpolator" at each station was just the sample mean across all stations; we should expect all the kriging methods to perform much better than this.

Results are shown in Fig. 4. The cross-validated mean squared error surfaces are very similar for each of the three kriging methods, but somewhat different for the sample mean interpolator. The overall cross-validated mean squared prediction errors (CVMSE) are .02151 for the sample mean interpolator, .01511 for the MLE-based kriging interpolator, .01522 for the SB-based kriging interpolator, .01515 for the HYB-based kriging interpolator. We can see that when assessed by CVMSE, the three kriging methods are almost identical and, as expected, substantially better than the sample mean interpolator.

Finally, we return to the question of whether the four subregions indicated in Fig. 1 are comparable, in the sense that the same spatial model (26) would fit to each. Of course, it would be possible to consider more than four subregions, or different definitions of the subregions, but since our primary purpose is to check on the overall stationarity of the model, we felt this comparison would be adequate for that.

Table 6 repeats the calculation of Table 5, but separately in each subregion. For one thing, this allows us to extend the comparison of parameter estimates using the MLE, SB and HYB methods. For example, combining Tables 5 and 6, there are 15 possible comparisons of parameter and region for which we can compare the three point estimates; in 11 of the 15 cases, the difference between HYB and MLE is smaller than the difference between SB and MLE. This confirms the impression created by the theoretical results, that overall HYB seems superior to SB, but it's not a totally one-sided comparison. We may also look at the standard errors of the three methods; in most cases, these are comparable.

As for the differences in parameter estimates among subregions, there is some evidence in Table 6 that they may be significant. This was further investigated by likelihood ratio tests; for each ordered pair of subregions, the MLE from the first subregion was inserted into a likelihood ratio test of the second subregion, to determine whether the parameter estimates were significantly different from the MLE for the second subregion. The resulting approximate χ^2_3 statistics ranged

Subregion	Method	α	ρ	ϕ
1	MLE	–4.83174 (0.46974)	0.77205 (0.49036)	0.49362 (0.15897)
1	SB	–5.02242 (0.70732)	0.32748 (0.24417)	0.51695 (0.32663)
1	HYB	–5.17059 (0.64551)	0.46043 (0.37098)	0.58790 (0.25137)
2	MLE	–4.38368 (0.33252)	0.33361 (0.14714)	0.15454 (0.23716)
2	SB	–4.52754 (0.35235)	0.40298 (0.21228)	0.22286 (0.24908)
2	HYB	–4.34438 (0.29689)	0.29654 (0.12450)	0.07623 (0.24124)
3	MLE	–4.24219 (0.39113)	0.56362 (0.27827)	0.45358 (0.15595)
3	SB	–4.60346 (0.47527)	1.66048 (1.91256)	0.61359 (0.14966)
3	HYB	–4.25600 (0.42109)	0.47405 (0.24954)	0.45278 (0.19369)
4	MLE	–4.40387 (0.39751)	0.62726 (0.38096)	0.46345 (0.14972)
4	SB	–4.45706 (0.37389)	0.65396 (0.38465)	0.48804 (0.15131)
4	HYB	–4.37124 (0.41044)	0.78797 (0.56558)	0.46578 (0.13622)

Table 6. Similar to Table 5, but estimated computed separately on each subregion.

from 1.5 to 38.2; in 8 of the possible 12 cases, there were statistically significant at the .05 level. Nevertheless, when applied to the cross-validated prediction errors, the prediction errors associated with global method (fitting a common model to all subregions) were comparable with those for a local method in which a separate model was fitted to each subregion as a preliminary to kriging. Only in subregion 2 was there a slightly larger discrepancy between the two sets of prediction errors (Fig. 5). This suggests that, while there may indeed be a discrepancy among the four subregions as measured by likelihood ratio tests for the parameter estimates, the practical effect of this on prediction is very slight, and so justifies our overall analysis based on a stationary model.

6 Summary and Conclusions

We began by defining three estimators, "big blocks", "small blocks" and "hybrid", all of which start by dividing a set of N observations into B blocks of approximate size K, where $BK \approx N$. Computational efficiency is maximized when B is between $N^{1/2}$ and $N^{2/3}$.

These estimators have been compared both in terms of their statistical efficiency, as measured by the ratio of the variance of our proposed estimator to that of the corresponding MLE, and by the IS/Direct ratio, which is the ratio of variance as estimated by the information sandwich formula

to the direct method based on the information matrix alone. If this ratio is much larger than 1, the variance of the estimator is seriously underestimated by using the direct method.

Comparisons of efficiencies suggest that the big blocks estimator is poor except when the range of the spatial covariance function is comparable with the range of the sampling locations. The efficiencies of the small blocks and hybrid estimators appear comparable in most circumstances, with the curious exception of the Matérn model with small shape parameter, when the hybrid method appears clearly superior.

In these models, the IS/Direct ratio is often close to 1 but in some cases is much bigger than 1. However, in all cases where the IS/Direct ratio was much bigger than 1, it was much smaller for the hybrid estimator than for the small blocks estimator. This provides another practical reason for using the hybrid estimator: even if it does not provide much advantage (over the small blocks estimator) in terms of statistical efficiency, it appears that the direct method of estimating variances is more satisfactory for the hybrid estimator than it is for the small blocks estimator.

A real-data example based on rainfall trends suggests that the hybrid estimator is very often, but not invariably, closer to the true MLE than the small-blocks estimator, while the three sets of standard errors (using the direct method) are comparable. The quality of predictions produced by the three methods, assessed by a cross-validated mean squared prediction error, was almost identical for this example.

In summary, the small-blocks and hybrid methods both appear to be good substitutes for the MLE across a very wide range of theoretical and practical examples. The big blocks method appears to be useful in practice only when the range of spatial covariances is comparable with the diameter of the region. As for the comparison between the small blocks and hybrid methods, in many cases the two are comparable in terms of efficiency, but the hybrid method is superior in the sense that estimated standard errors from inverting the approximate observed information matrix are closer to the true standard errors than those derived by the small blocks method. We therefore recommend the hybrid method as a good all-round alternative to exact maximum likelihood.

7 References

Aho, A.V., Hopcroft, J.E. and Ullmann, J.D. (1974), *The Design and Analysis of Computer Algorithms*. Addison Wesley.

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2003), *Hierarchichal Modeling and Analysis for*

Spatial Data. CRC Press/Chapman and Hall, Boca Raton, FL.

Caragea, P. and Smith, R.L. (2005), Asymptotic properties of three alternative estimators of spatial parameters. In preparation.

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley, New York.

Cressie, N. (1993), *Statistics for Spatial Data*. John Wiley, New York.

Diggle, P. J. and Tawn, J. A. and Moyeed, R. A. (1998), Model-based geostatistics. *Applied Statistics* **47**, 299–326.

Fuentes, M. and Smith R.L. (2001), A new class of nonstationary spatial models. North Carolina State University and University of North Carolina Institute of Statistics Mimeo Series #2534. http://www4.stat.ncsu.edu/ fuentes/nonstat.ps

Guyon, X. (1982), Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika* **69**, 95–105.

Heyde, C.C, (1997), *Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation*. Springer, New York.

Higdon, D., Swall, J. and Kern, J. (1999), Non-stationary spatial modeling. In *Bayesian Statistics 6*, eds. J.M. Bernardo *et al.*, Oxford University Press, pp. 761–768.

Journel, A.G. and Huijbregts, C.J. (1978), *Mining Geostatistics*. Academic Press, London.

Karl, T.R. and Knight, R.W. (1998), Secular trends of precipitation amount, frequency, and intensity in the USA. *Bull. Amer. Meteor. Soc.* **79**, 231–241.

Kitanidis, P.K. (1983), Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research* **19**, 909–921.

Mardia, K.V. and Marshall, R.J. (1984), Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135-146.

Mardia, K.V. and Watkins, A.J. (1989), On multimodality of the likelihood in the spatial linear model. *Biometrika* **76**, 289–295.

Ripley, B.D. (1981), *Spatial Statistics*. Wiley, New York.

Ripley, B.D. (1988), *Statistical Inference for Spatial Processes*. Cambridge University press, Cambridge, U.K.

Sampson, P.D. and Guttorp, P. (1992), Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87**, 108-119.

Smith, R.L. (2001), *Environmental Statistics*, course notes.
http://www.unc.edu/depts/statistics/postscript/rs/envnotes.pdf

Stein, M.L. (1999), *Interpolation of Spatial Data: Some Theory of Kriging.* Springer Verlag, New York.

Stein, M.L., Chi, Z. and Welty, L.J. (2004), Approximating likelihoods for large spatial data sets. *J.R. Statist.Soc. B* **66**, 275–296.

Vecchia, A. V. (1988), Estimation and identification for continuous spatial processes. *J. Roy. Statist B* **50** 297–312.

Warnes, J.J. and Ripley, B.D. (1987), Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika* **74**, 640–642.

Whittle, P. (1954), On stationary processes in the plane. *Biometrika* **41**, 434–439.

Zimmerman, D.L. (1989), Computationally exploitable structure of covariance matrices and generalized covariance matrices in spatial models. *J. Statist. Comput. Simul.* **32**, 1–15.

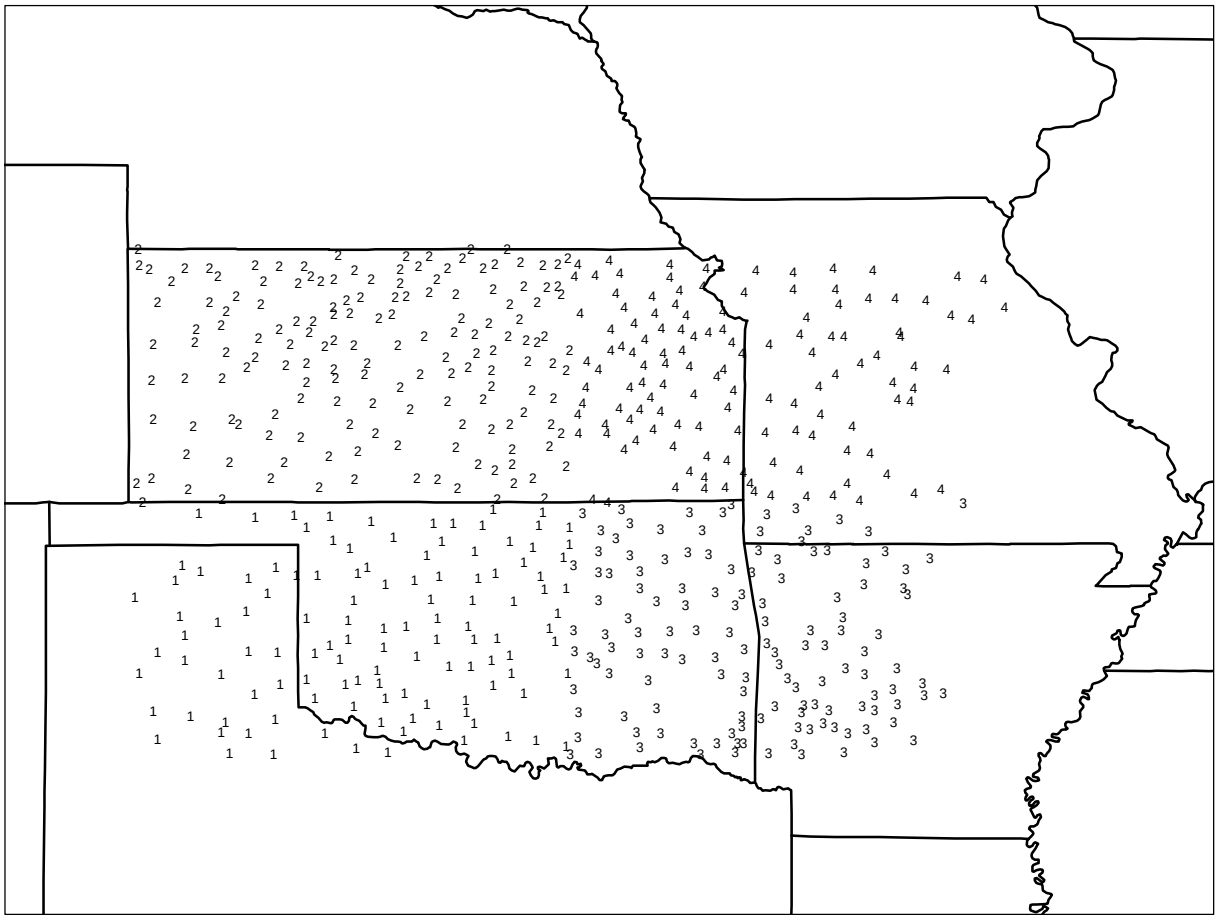

Figure 1. The study region; 540 stations divided into four subregions.

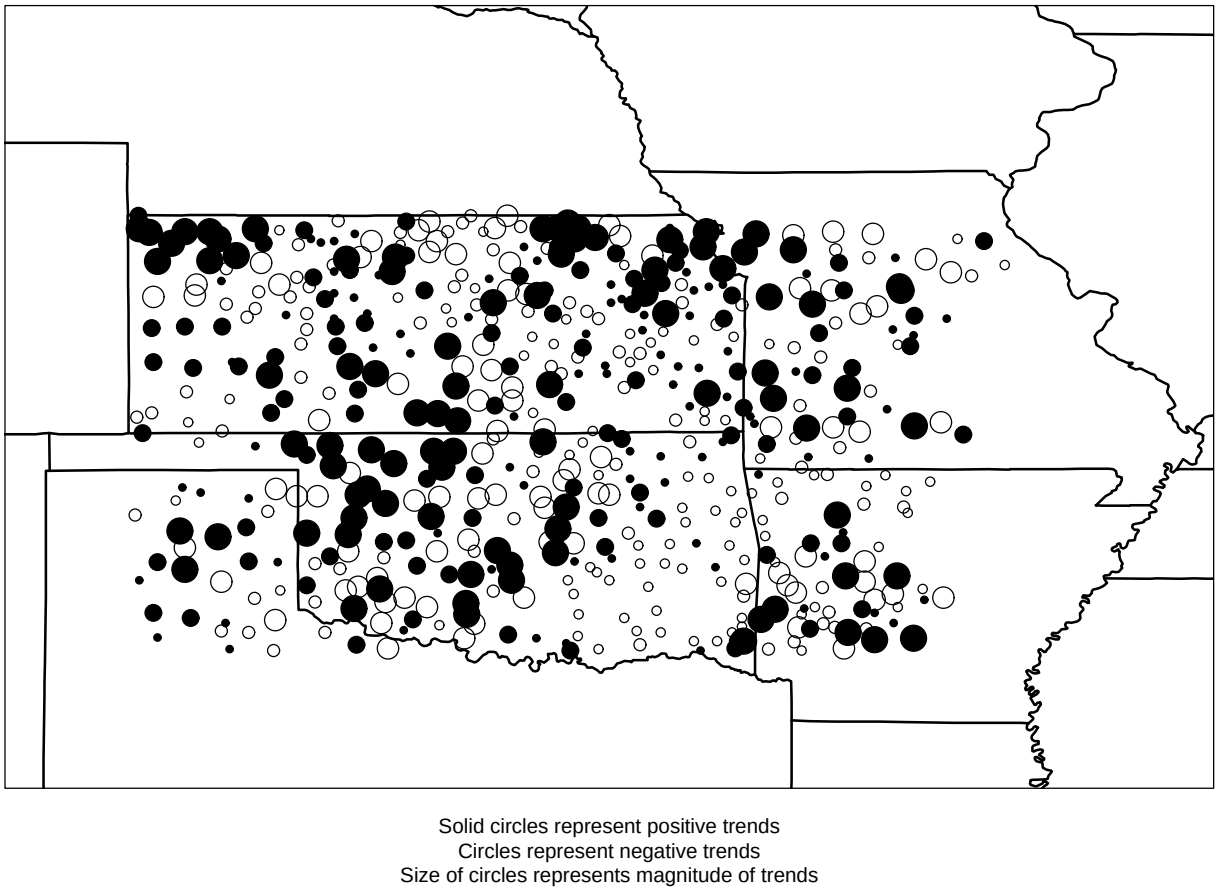

Figure 2. Estimated linear trends in rainfall monthly totals, 1966–1984, fitted pointwise to each station.

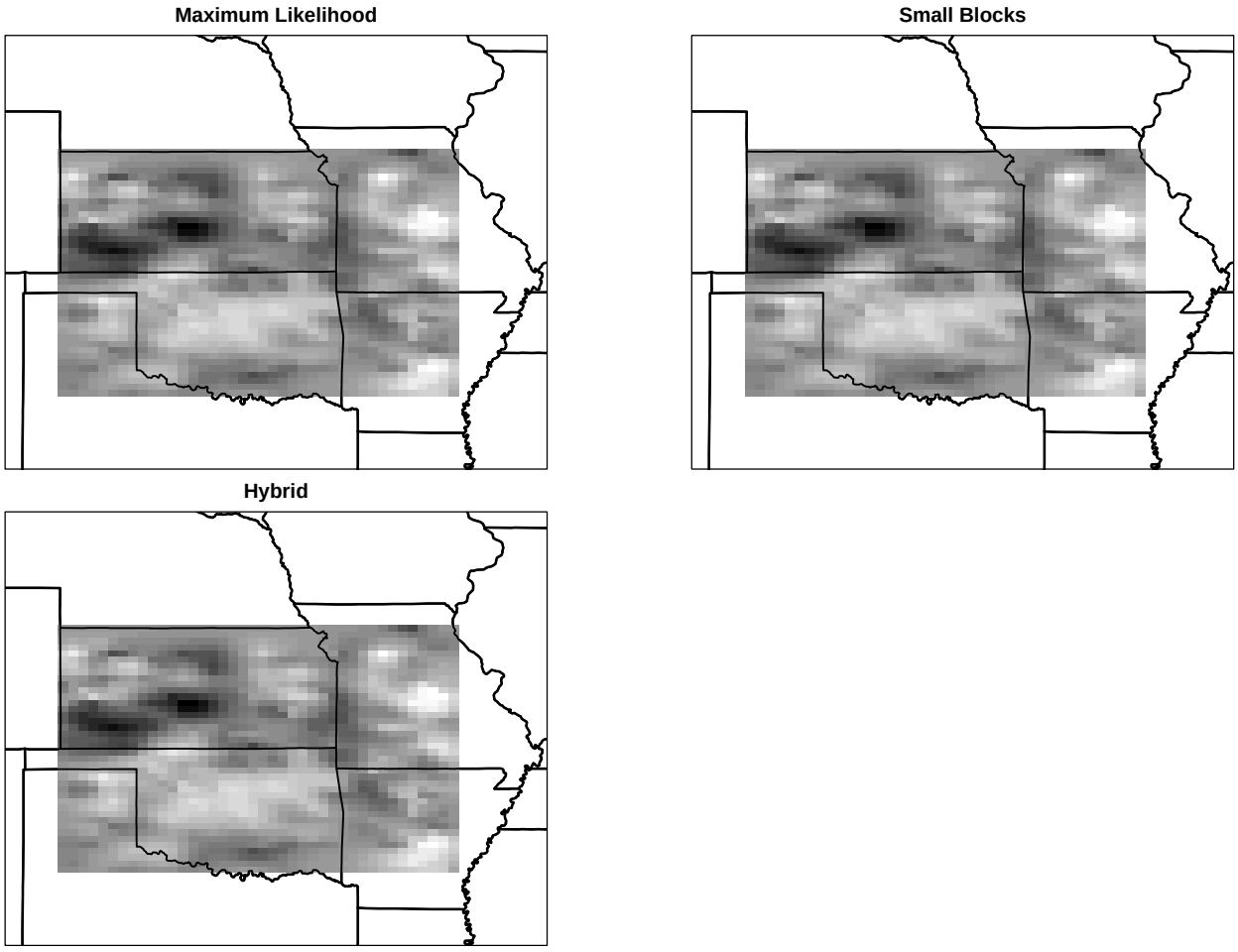

Figure 3. Interpolated trend fitted using exact maximum likelihood as well as the small blocks and hybrid approaches; exponential covariance model with nugget.

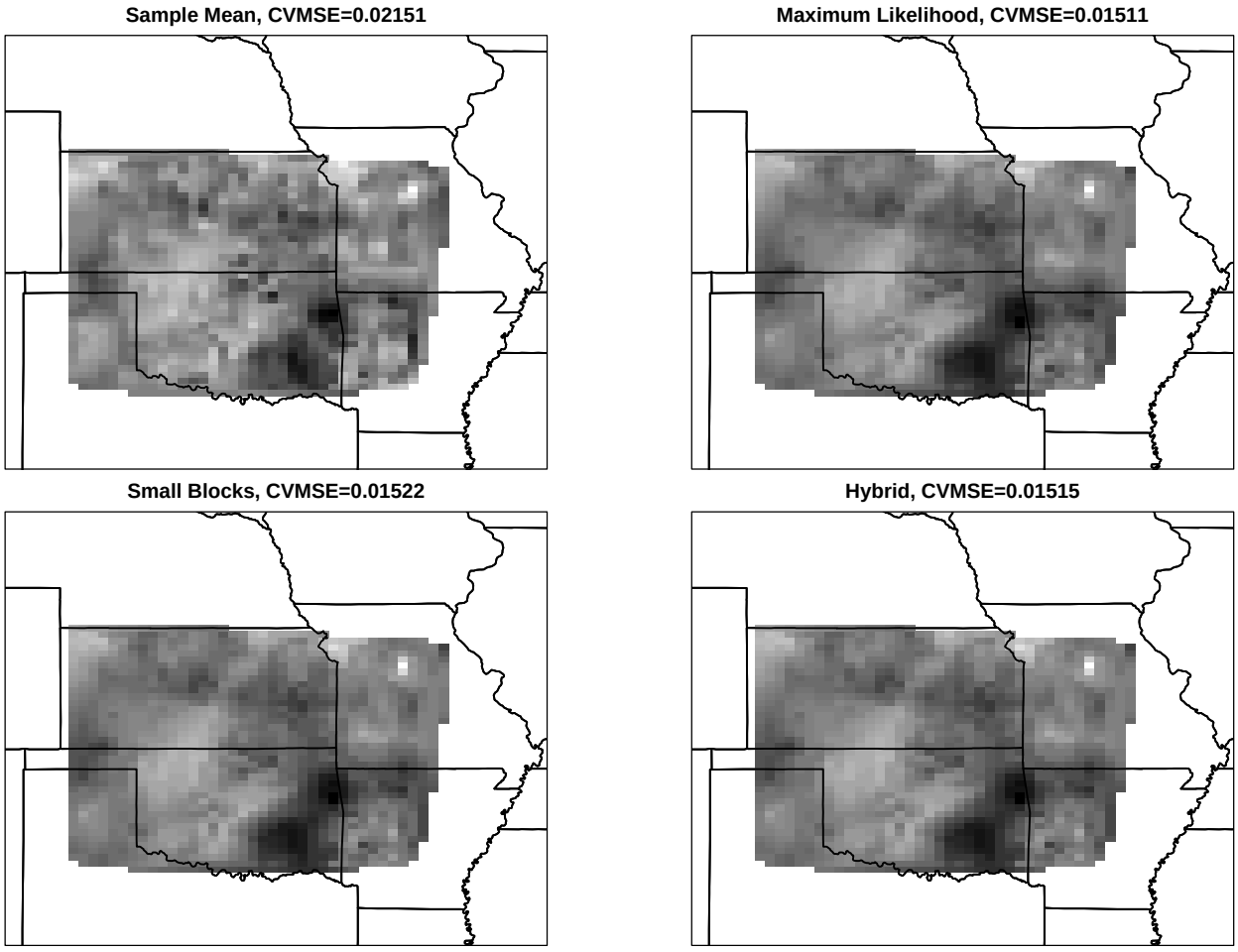

Figure 4. Cross-validated mean squared prediction errors for the sample-mean predictor and for kriging using each of the three estimators.

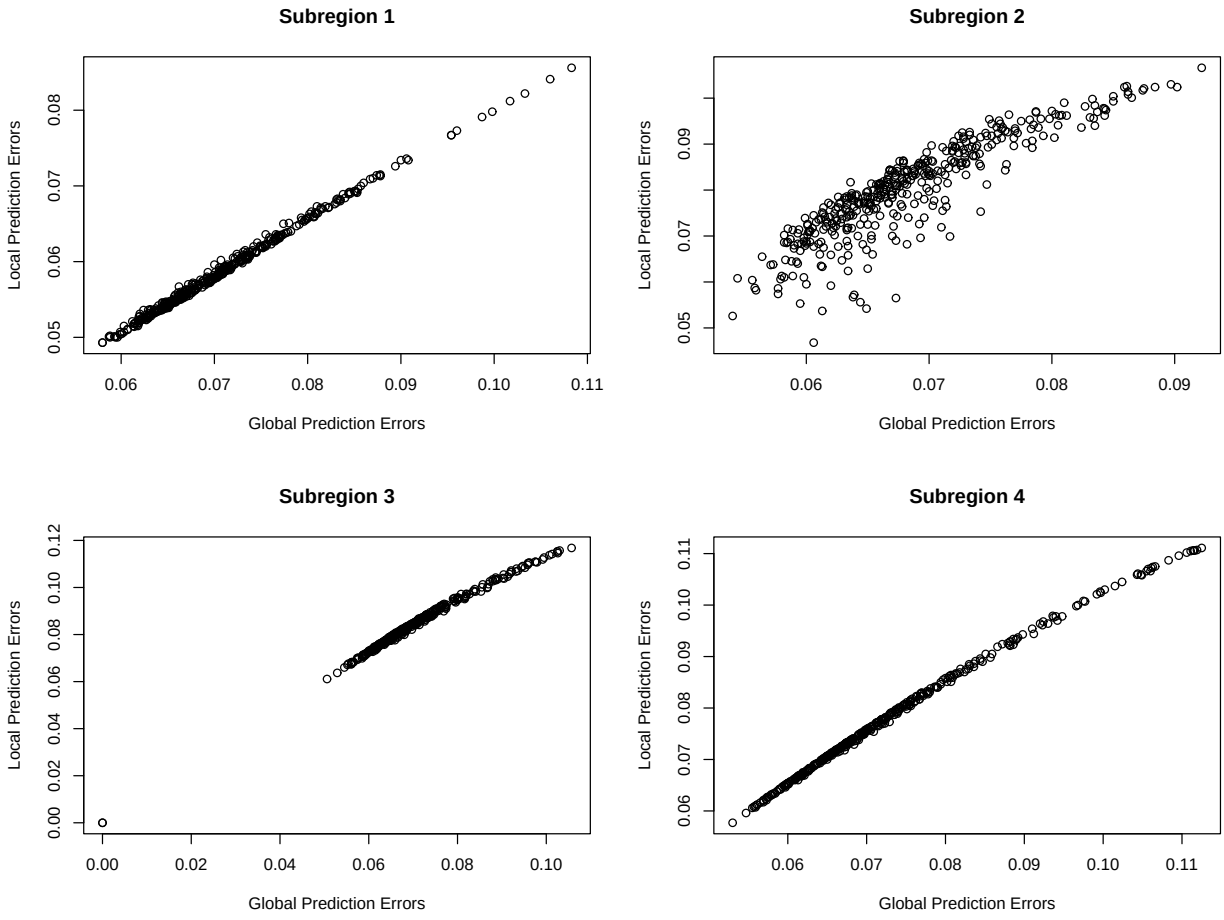

Figure 5. Comparison of local and global prediction errors for each subregion. Global prediction error based on model (26) fitted to all four subregions; local prediction error based on separate model for each subregion.