Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Approximate Bayesian inference for large spatial datasets using predictive process models

Jo Eidsvik[a,*], Andrew O. Finley[b], Sudipto Banerjee[c], Håvard Rue[a]

[a] *Department of Mathematical Sciences, NTNU, Trondheim, Norway*

[b] *Department of Forestry, Michigan State University, MI, USA*

[c] *Department of Biostatistics, University of Minnesota, MN, USA*

ARTICLE INFO

Article history:
Received 21 December 2010
Received in revised form 24 October 2011
Accepted 28 October 2011
Available online xxxx

Keywords:
Approximate Bayesian inference
Computational statistics
Gaussian processes
Geostatistics
Laplace approximation
Predictive process model

ABSTRACT

The challenges of estimating hierarchical spatial models to large datasets are addressed. With the increasing availability of geocoded scientific data, hierarchical models involving spatial processes have become a popular method for carrying out spatial inference. Such models are customarily estimated using Markov chain Monte Carlo algorithms that, while immensely flexible, can become prohibitively expensive. In particular, fitting hierarchical spatial models often involves expensive decompositions of dense matrices whose computational complexity increases in cubic order with the number of spatial locations. Such matrix computations are required in each iteration of the Markov chain Monte Carlo algorithm, rendering them infeasible for large spatial datasets. The computational challenges in analyzing large spatial datasets are considered by merging two recent developments. First, the predictive process model is used as a reduced-rank spatial process, to diminish the dimensionality of the model. Then a computational framework is developed for estimating predictive process models using the integrated nested Laplace approximation. The settings where the first stage likelihood is Gaussian or non-Gaussian are discussed. Issues such as predictions and model comparisons are also discussed. Results are presented for synthetic data and several environmental datasets.

1. Introduction

The advent of geographic information systems have led to accurate geocoding of locations where massive amounts of scientific data are collected. This has generated considerable interest in statistical modeling for such data; see, for example, the books by Cressie (1993), Banerjee et al. (2004), and Schabenberger and Gotway (2004). Here, we focus upon the setting where the number of locations yielding observations is too large for fitting desired hierarchical spatial random effects models. Full inference and accurate assessment of uncertainty involves matrix decompositions whose complexity increases as $O(n^3)$ in the number of locations, n, hence the infeasibility or "big n" problem for large datasets.

Modeling large spatial datasets has received much attention in the recent past. Vecchia (1988) proposed approximating the likelihood with a product of appropriate conditional distributions to obtain maximum-likelihood estimates. Stein et al. (2004) adapt this to restricted maximum likelihood estimation. Another possibility is to approximate the likelihood using spectral representations of the spatial process like Fuentes (2007). These likelihood approximations yield a joint distribution, but not a process that facilitates spatial interpolation. Yet another approach considers compactly supported correlation functions, see e.g. Furrer et al. (2006), Kaufman et al. (2008), Du et al. (2009) and Sang and Huang (2012), that yield

[*] Corresponding author. Tel.: +47 90127472; fax: +47 73593524.
E-mail address: joeid@math.ntnu.no (J. Eidsvik).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.10.022

sparse correlation structures. More efficient sparse solvers can then be employed for kriging and variance estimation, but the tapered structures may limit modeling flexibility. Also, full likelihood-based inference still requires determinant computations that may be problematic.

Rather than approximations, one could build models especially geared towards handling of large spatial datasets. These are representations of the spatial process in a lower-dimensional subspace and are often referred to as low-rank or reduced-rank spatial models, see Higdon (2002), Kammann and Wand (2003), Stein (2007, 2008), Cressie and Johannesson (2008), Banerjee et al. (2008) and Crainiceaniu et al. (2008). Many of these methods are variants of the so-called "subset of regressors" methods used in Gaussian process regressions for large datasets in machine learning, e.g. Rasmussen and Williams (2006). The idea here is to consider a smaller set of locations, or "knots", say $\mathcal{S}^* = \{\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_{n^*}^*\}$, where the number of knots, n^*, is *fixed* to be much smaller than the number of observed sites, and to express the spatial process realizations over n locations in terms of its realizations over the smaller set of knots. It is reasonable to assume there will be insignificant loss of spatial information in the underlying process from using a smaller set of locations – the knots – with adequate domain coverage. Subsequently, we will consider a special class of low-rank processes called the *predictive process*, see Banerjee et al. (2008). This arises from a conditional expectation of the original process (often referred to as the *parent process*) given its realization over the knots. As such, the predictive process model is a dimension reduction technique that requires no additional tuning parameters in the modeling.

A key issue in predictive process modeling is the number and selection of knots, which is a challenging problem, with choice in two dimensions more difficult than in one. The choice of n^* is governed by computational cost and sensitivity to choice. Customarily, the analysis is implemented over different choices of n^* and knot locations. The issue is not dissimilar to a spatial design problem, e.g. Nychka and Saltzman (1998), Xia et al. (2006) and Diggle and Lophaven (2006). The standard method is to experiment with different knot configurations. Using Markov chain Monte Carlo (MCMC) for such experimentations will, however, be a daunting task and fast, accurate approximation methods will need to be explored.

In recent work Rue et al. (2009) propose an Integrated Nested Laplace Approximation (INLA) algorithm as an alternative to MCMC for latent Gaussian models. INLA presents a very versatile template for estimating latent Gaussian models by repeated use of the Laplace Approximation (LA), see Tierney and Kadane (1986). Rue et al. (2009) use computationally effective Gaussian Markov random field approximations, see Rue and Held (2005), to deliver fast and accurate approximations to posterior marginals. Eidsvik et al. (2009) use the same Laplace techniques for irregular moderate size data from a spatial Generalized Linear Mixed Model (GLMM). Extensive studies conducted by Eidsvik et al. (2009) and Rue et al. (2009) reveal that, for a wide class of latent Gaussian models, INLA produces inference that is essentially indistinguishable from MCMC in a mere fraction of the time required by the latter. The key to successful use of INLA, is a reasonable Gaussian approximation to the full conditional of the latent variables, including regression effects. A numerical optimization and integration routine is used for the covariance hyperparameters. The LA has been a powerful tool in statistical inference. Frequentist approaches use the LA for marginalized likelihood inference, see e.g. Breslow and Clayton (1993), Ainsworth and Dean (2006) and Evangelou et al. (2011). In the Bayesian context it has been applied for model choice using Bayes factors, but then the full conditionals are usually approximated by sampling, see e.g. Chib (1995) and Lewis and Raftery (1997). Hsiao et al. (2004) use the LA for related purposes.

This article presents a framework for estimating predictive process models using INLA. The remainder of the article evolves as follows. Section 2 discusses the spatial predictive process, its properties and how it is employed in hierarchical spatial GLMM context. Section 3 outlines approximate Bayesian inference using INLA. Section 4 considers a number of simulation experiments as well as practical illustrations from fisheries and forestry. Finally, Section 5 concludes the article with a discussion and an eye towards future work.

2. Hierarchical modeling with the predictive process

In this section we will present the predictive process models for Gaussian processes and for GLMMs. Our exposition is meant to facilitate the use of approximate Bayes inference methods applied to these models in Section 3.

2.1. The Gaussian predictive process

Geostatistical settings typically assume, at locations $\boldsymbol{s} \in D \subseteq \Re^2$, a Gaussian response variable $Y(\boldsymbol{s})$ along with a $p \times 1$ vector of spatially referenced predictors $\boldsymbol{x}(\boldsymbol{s})$ which are associated through a spatial regression model such as,

$$Y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s}). \tag{1}$$

That is, the residual comprises a spatial process, $w(\boldsymbol{s})$, and an independent process, $\epsilon(\boldsymbol{s})$, often called the *nugget*. The $w(\boldsymbol{s})$ are spatial random effects, providing local adjustment (with structured dependence) to the mean, interpreted as capturing the effect of unmeasured or unobserved covariates with spatial pattern.

The customary process specification for $w(\boldsymbol{s})$ is a mean 0 Gaussian process with covariance function, $C(\boldsymbol{s}_1, \boldsymbol{s}_2)$, denoted $GP(0, C(\boldsymbol{s}_1, \boldsymbol{s}_2))$. We often specify $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \sigma^2\rho(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\phi})$ where $\rho(\cdot; \boldsymbol{\phi})$ is a correlation function and $\boldsymbol{\phi}$ includes spatial decay and smoothness parameters, yielding a constant process variance. In any event, $\epsilon(\boldsymbol{s}) \overset{iid}{\sim} N(0, \tau^2)$ for every location $\boldsymbol{s}$. Prior distributions on the remaining parameters complete the hierarchical model. Customarily, the regression effect $\boldsymbol{\beta}$

is assigned a multivariate Gaussian prior, i.e. $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_\beta, \Sigma_\beta)$, while the latent variance component σ^2 and the nugget variance τ^2 are assigned $IG(\cdot,\cdot)$ priors. The process correlation parameter(s), $\boldsymbol{\phi}$, are usually assigned some informative priors (e.g. uniform over a finite range) based upon the underlying spatial domain.

With n locations, say $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, the process realizations are collected into an $n \times 1$ vector, say $\boldsymbol{w} = (w(\boldsymbol{s}_1), \ldots, w(\boldsymbol{s}_n))'$, which follows a multivariate normal distribution with mean $\mathbf{0}$ and dispersion matrix $\sigma^2\boldsymbol{R}(\boldsymbol{\phi})$ with $\rho(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\phi})$ being the (i,j)-th element of $\boldsymbol{R}(\boldsymbol{\phi})$. Letting $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))'$ be the $n \times 1$ vector of observed responses, we obtain a Gaussian likelihood that combines with the customary hierarchical specifications to yield a posterior distribution

$$\pi(\boldsymbol{\beta}, \boldsymbol{w}, \sigma^2, \tau^2, \boldsymbol{\phi} \mid \boldsymbol{Y}) \propto \pi(\boldsymbol{\phi}) \times IG(\tau^2 \mid a_\tau, b_\tau) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \Sigma_\beta) \\ \times N(\boldsymbol{w} \mid \mathbf{0}, \sigma^2\boldsymbol{R}(\boldsymbol{\phi})) \times \prod_{i=1}^{n} N(Y(\boldsymbol{s}_i) \mid \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + w(\boldsymbol{s}_i), \tau^2). \tag{2}$$

Often a marginalized likelihood is used that is obtained by integrating out the spatial effects $\boldsymbol{w}$ and the regression coefficients $\boldsymbol{\beta}$. This yields

$$\pi(\sigma^2, \tau^2, \boldsymbol{\phi} \mid \boldsymbol{Y}) \propto \pi(\boldsymbol{\phi}) \times IG(\tau^2 \mid a_\tau, b_\tau) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) N\left(\boldsymbol{Y} \mid \boldsymbol{X}\boldsymbol{\mu}_\beta, \boldsymbol{X}\boldsymbol{\Sigma}_\beta\boldsymbol{X}' + \sigma^2\boldsymbol{R}(\boldsymbol{\phi}) + \tau^2\boldsymbol{I}\right), \tag{3}$$

where row i of matrix $\boldsymbol{X}$ is $\boldsymbol{x}(\boldsymbol{s}_i)'$. This marginalization over $\boldsymbol{w}$ and $\boldsymbol{\beta}$ can be interpreted as a ratio of joints and conditionals since,

$$\pi(\boldsymbol{Y} \mid \cdot) = \int_{\boldsymbol{w},\boldsymbol{\beta}} \pi(\boldsymbol{Y}, \boldsymbol{w}, \boldsymbol{\beta} \mid \cdot) d\boldsymbol{w} d\boldsymbol{\beta} = \frac{\pi(\boldsymbol{Y}, \boldsymbol{w}, \boldsymbol{\beta} \mid \cdot)}{\pi(\boldsymbol{w}, \boldsymbol{\beta} \mid \boldsymbol{Y}, \cdot)}. \tag{4}$$

In fact, we will utilize this in the LA below. The marginal posterior distribution of the spatial effects and regression parameters is given by

$$\pi(\boldsymbol{w}, \boldsymbol{\beta} \mid \boldsymbol{Y}) = \int \pi\left(\boldsymbol{w}, \boldsymbol{\beta} \mid \boldsymbol{Y}, \sigma^2, \tau^2, \boldsymbol{\phi}\right) \pi(\sigma^2, \tau^2, \boldsymbol{\phi} \mid \boldsymbol{Y}) d\sigma^2 d\tau^2 d\boldsymbol{\phi},$$

where $\pi\left(\boldsymbol{w}, \boldsymbol{\beta} \mid \boldsymbol{Y}, \sigma^2, \tau^2, \boldsymbol{\phi}\right)$ is a multivariate normal distribution.

Irrespective of whether we use (2) or (3), estimation and prediction will require matrix factorizations involving the dense $n \times n$ matrix $\boldsymbol{R}(\boldsymbol{\phi})$ which may become prohibitively expensive for large n. Recently Banerjee et al. (2008) proposed a class of knot-based spatial process models for large spatial datasets. These models consider a fixed set of n^* knots, with $n^* \ll n$, which may or may not form a subset of the entire collection of observed locations. The Gaussian process $w(\boldsymbol{s})$ yields an n^*-vector of realizations over the knots, say $\boldsymbol{w}^* = (w(\boldsymbol{s}_1^*), \ldots, w(\boldsymbol{s}_{n^*}^*))'$, which follows a $N\{\mathbf{0}, \sigma^2\boldsymbol{R}^*(\boldsymbol{\phi})\}$, where $\boldsymbol{R}^*(\boldsymbol{\phi}) = \{\rho(\boldsymbol{s}_i^*, \boldsymbol{s}_j^*; \boldsymbol{\phi})\}_{i,j=1}^{n^*}$ is the corresponding $n^* \times n^*$ dispersion matrix. Spatial interpolation (or "kriging") at a generic site $\boldsymbol{s}$ is executed through

$$\tilde{w}(\boldsymbol{s}) = E\{w(\boldsymbol{s}) \mid \boldsymbol{w}^*\} = \boldsymbol{r}(\boldsymbol{s}; \boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})\boldsymbol{w}^*. \tag{5}$$

This yields a spatial process $\tilde{w}(\boldsymbol{s}) \sim GP\{0, \sigma^2\tilde{\rho}(\cdot)\}$ where $\tilde{\rho}(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\phi}) = \boldsymbol{r}(\boldsymbol{s}_1; \boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})\boldsymbol{r}(\boldsymbol{s}_2, \boldsymbol{\phi})$ and $\boldsymbol{r}(\boldsymbol{s}; \boldsymbol{\phi})$ is the $n^* \times 1$ vector whose j-th element is given by $\rho(\boldsymbol{s}, \boldsymbol{s}_j^*; \boldsymbol{\phi})$. We refer to $\tilde{w}(\boldsymbol{s})$ as the *predictive process* derived from the *parent process* $w(\boldsymbol{s})$. The predictive process is a spatially adaptive linear transformation of the realizations of $w(\boldsymbol{s})$ over the knot locations $\mathcal{S}^*$, with $\boldsymbol{r}(\boldsymbol{s}; \boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})$ comprising the coefficients of the transformation. This also implies that $\tilde{w}(\boldsymbol{s})$ is non-stationary, even though $w(\boldsymbol{s})$ is not.

Replacing $w(\boldsymbol{s})$ in (1) with $\tilde{w}(\boldsymbol{s})$, we obtain the predictive process model,

$$Y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + \tilde{w}(\boldsymbol{s}) + \epsilon(\boldsymbol{s}). \tag{6}$$

Using (6) as the likelihood, we obtain the predictive process counterpart of (2) as

$$\pi(\boldsymbol{\phi}) \times IG(\tau^2 \mid a_\tau, b_\tau) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \Sigma_\beta) \\ \times N(\boldsymbol{w}^* \mid \mathbf{0}, \sigma^2\boldsymbol{R}^*(\boldsymbol{\phi})) \times \prod_{i=1}^{n} N(Y(\boldsymbol{s}_i) \mid \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \tilde{w}(\boldsymbol{s}_i), \tau^2). \tag{7}$$

Dimension reduction occurs since the computations now involve evaluating the $n^* \times n^*$ matrix $\boldsymbol{R}^{*-1}(\boldsymbol{\phi})$, where n^* is chosen to be much smaller than n. The method trades computer time for a required selection of knots. Unlike other knot-based methods, the predictive process neither introduces any additional parameters nor involves projecting data onto a grid, while enjoying attractive theoretical properties that justify its use as a *best approximation* for the parent process. For example, $\tilde{w}(\boldsymbol{s})$ is an orthogonal projection of $w(\boldsymbol{s})$ on an appropriate linear subspace, e.g. Stein (1999), minimizing $E[\{w(\boldsymbol{s}) - f(\boldsymbol{w}^*)\}^2 \mid \boldsymbol{w}^*]$ over all real-valued functions $f(\boldsymbol{w}^*)$.

Rather than an approximation to the parent process, we consider the predictive process as a dimension-reducing model for large point-referenced datasets. It is crucial, therefore, that its parameters should be interpreted with respect to (6) and not (1). In fact, being smoother than the parent process, the predictive process tends to have lower variance which, in turn, leads to an upward bias in the nugget. The following inequality reflects, more formally, the shrinkage in variability for the predictive process

$$\mathrm{var}\{w(\boldsymbol{s})\} = \mathrm{var}\{\mathrm{E}[w(\boldsymbol{s}) \mid \boldsymbol{w}^*]\} + \mathrm{E}\{\mathrm{var}[w(\boldsymbol{s}) \mid \boldsymbol{w}^*]\} \geq \mathrm{var}\{\mathrm{E}[w(\boldsymbol{s}) \mid \boldsymbol{w}^*]\} = \mathrm{var}\{\tilde{w}(\boldsymbol{s})\}.$$

The diminished variability in $\tilde{w}(\boldsymbol{s})$ is often manifested by an overestimation of the nugget variance τ^2. Banerjee et al. (2010) explore these biases in greater detail.

Finley et al. (2009a) consider modifying the predictive process by adding a heteroscedastic white-noise Gaussian process. More specifically, they propose replacing $\tilde{w}(\boldsymbol{s})$ in (6) with $\tilde{w}_\epsilon(\boldsymbol{s}) = \tilde{w}(\boldsymbol{s}) + \tilde{\epsilon}(\boldsymbol{s})$, where $\tilde{\epsilon}(\boldsymbol{s}) \overset{iid}{\sim} N\left\{0, \sigma^2(1 - \boldsymbol{r}(\boldsymbol{s}; \boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})\boldsymbol{r}(\boldsymbol{s}; \boldsymbol{\phi}))\right\}$. The white-noise term can be regarded as an additional nugget effect in the likelihood model, i.e. $N\left\{Y(\boldsymbol{s}_i) \mid \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \tilde{w}(\boldsymbol{s}_i), \tau^2 + \sigma^2(1 - \boldsymbol{r}(\boldsymbol{s}; \boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})\boldsymbol{r}(\boldsymbol{s}; \boldsymbol{\phi}))\right\}$. Thus, adjusting for scale is straightforward in the predictive process. One could imagine a similar compensation for the correlation parameters $\boldsymbol{\phi}$, but in general it seems both theoretically and computationally harder to adjust for this parameter. Still, the predictive process model has provided reliable results for complex covariance structures, see e.g. Banerjee et al. (2008) and Finley et al. (2009a).

Let $\boldsymbol{v}^* = (\boldsymbol{w}^{*'}, \boldsymbol{\beta}')'$ be the $(n^* + p) \times 1$ vector collecting all a priori Gaussian effects. The likelihood for the modified predictive process is

$$N(\boldsymbol{Y} \mid \boldsymbol{H}^*\boldsymbol{v}^*, \sigma^2\boldsymbol{R}_{\tilde{\epsilon}} + \tau^2\boldsymbol{I}_n), \quad \boldsymbol{H}^* = [\mathcal{F}(\boldsymbol{\phi}), \boldsymbol{X}], \tag{8}$$

where $\mathcal{F}(\boldsymbol{\phi}) = \mathcal{R}(\boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})$, and $\mathcal{R}(\boldsymbol{\phi})'$ is the $n \times n^*$ matrix whose i-th row is given by $\boldsymbol{r}(\boldsymbol{s}_i; \boldsymbol{\phi})'$, for $i = 1, \ldots, n$. Further, $\boldsymbol{R}_{\tilde{\epsilon}}$ is the modification part written as an $n \times n$ diagonal matrix with i-th diagonal element $\left\{1 - \boldsymbol{r}(\boldsymbol{s}_i; \boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})\boldsymbol{r}(\boldsymbol{s}_i; \boldsymbol{\phi})\right\}$. An expression for the marginalized posterior distribution, the modified predictive process counterpart to (3), can be obtained by integrating out $\boldsymbol{v}^*$, whereupon we have

$$\pi(\boldsymbol{\phi}) \times IG(\tau^2 \mid a_\tau, b_\tau) \times IG(\sigma^2 \mid a_\sigma, b_\sigma)N(\boldsymbol{Y} \mid \boldsymbol{X}\boldsymbol{\mu}_\beta, \boldsymbol{X}\Sigma_\beta\boldsymbol{X}' + \sigma^2\mathcal{F}(\boldsymbol{\phi})\boldsymbol{R}^*(\boldsymbol{\phi})\mathcal{F}(\boldsymbol{\phi})' + \sigma^2\boldsymbol{R}_{\tilde{\epsilon}} + \tau^2\boldsymbol{I}_n). \tag{9}$$

The number of knots n^* and selection of the sites that will act as knots is a complex problem and raises the question of whether to use a subset of the observed spatial locations or a disjoint set of locations. Finley et al. (2009a) explored the knot selection issue for predictive processes. Key guidelines known from the spline literature are to place knots such that one covers the domain and have more knots where data are dense. The prediction gets worse away from knots, and the optimal knot placement (based on some criterion) would in general depend on the spatial correlation. On average, one would improve the predictions by adding knots sequentially, but there are no directly available bounds for the approximation error. In practice, it is feasible to estimate predictive process models for various n^* and with different choices of knots to arrive at configurations yielding reliable and robust inference.

2.2. *Predictive process models with non-Gaussian likelihoods*

We now consider the setting with non-Gaussian likelihoods. There are two typical non-Gaussian GLMM first stage settings: (i) binary response at locations modeled using logit or probit regression and (ii) count data at locations modeled using Poisson regression. Diggle et al. (1998) unify the use of these GLMMs in spatial data contexts. See also Lin et al. (2000), Kammann and Wand (2003), Banerjee et al. (2004), Finley et al. (2009b) and Latimer et al. (2009). Essentially, we construct the likelihood assuming conditional independence of the outcomes, i.e. the $Y(\boldsymbol{s}_i)$'s, which arise from an exponential family. In other words, we replace (1) with the assumption that the expected value is linear on a transformed scale, i.e., $\eta(\boldsymbol{s}) \equiv g(E(Y(\boldsymbol{s}))) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + w(\boldsymbol{s})$, where $g(\cdot)$ is a suitable link function. More specifically, the resulting posterior would take a form analogous to (2):

$$\pi(\boldsymbol{\phi}) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \Sigma_\beta)N(\boldsymbol{w} \mid \boldsymbol{0}, \sigma^2\boldsymbol{R}(\phi)) \times \prod_{i=1}^{n} \pi(Y(\boldsymbol{s}_i) \mid \eta(\boldsymbol{s}_i)), \tag{10}$$

where $\pi(Y(\boldsymbol{s}_i) \mid \eta(\boldsymbol{s}_i))$ belongs to the exponential family of densities.

For large datasets, we insert the predictive process, $\tilde{w}(\boldsymbol{s})$, in the link function so that $\eta(\boldsymbol{s}_i) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \tilde{w}(\boldsymbol{s}_i)$. Let again $\boldsymbol{v}^* = (\boldsymbol{w}^{*'}, \boldsymbol{\beta}')'$ be the $(n^* + p) \times 1$ vector comprising the realizations of the spatial predictive process and the regression parameters. The posterior is

$$\pi(\boldsymbol{v}^*, \sigma, \boldsymbol{\phi} \mid \boldsymbol{Y}) \propto \pi(\boldsymbol{\phi}) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times N(\boldsymbol{v}^* \mid \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) \prod_{i=1}^{n} \pi(Y(\boldsymbol{s}_i) \mid \boldsymbol{r}(\boldsymbol{s}_i; \boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})\boldsymbol{w}^* + \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta}), \tag{11}$$

where mean vector $\boldsymbol{\mu}^* = (\boldsymbol{0}_{n^*}, \boldsymbol{\mu}_\beta)$ and the $(n^* + p) \times (n^* + p)$ covariance matrix

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} \sigma^2\boldsymbol{R}^*(\boldsymbol{\phi}) & \boldsymbol{0}_{n^* \times p} \\ \boldsymbol{0}_{p \times n^*} & \boldsymbol{\Sigma}_\beta \end{bmatrix}. \tag{12}$$

The canonical length n parameter vector in the GLM likelihood model can be defined by $\boldsymbol{\eta} = \boldsymbol{H}^*\boldsymbol{v}^*$, where $\boldsymbol{H}^* = [\mathcal{F}(\boldsymbol{\phi}), \boldsymbol{X}]$, like in Section 2.1. Unlike with Gaussian likelihoods, analytical marginalization over the spatial and regression effects is no longer possible.

The modified predictive process can again be included by introducing $\tilde{w}_\epsilon(\boldsymbol{s}) = \tilde{w}(\boldsymbol{s}) + \tilde{\epsilon}(\boldsymbol{s})$ in the model. The additive, heteroscedastic, noise term $\tilde{\epsilon}$ cannot be directly added as an extra nugget term in this non-Gaussian likelihood situation. Instead, we can include it as a latent variable, giving an augmented state $\boldsymbol{v}^* = (\boldsymbol{w}^{*\prime}, \boldsymbol{\beta}', \tilde{\epsilon}')'$, and an associated $\boldsymbol{H}^* = [\mathcal{F}(\boldsymbol{\phi}), \boldsymbol{X}, \boldsymbol{I}_n]$. A priori, the noise term is conditionally independent of $\boldsymbol{\beta}$ and $\boldsymbol{v}^*$. The dimension of $\tilde{\epsilon}$ equals n, but the effects are independent (diagonal covariance), and one utilizes this in the evaluation.

General settings can be treated using these ideas with the appropriate choice of an exponential family member and a link function. For instance, with binomial data, $\pi(Y(\boldsymbol{s}_i) \mid \eta(\boldsymbol{s}_i)) \sim \mathit{Binomial}(N(\boldsymbol{s}_i), p(\eta(\boldsymbol{s}_i)))$, where $p(\eta(\boldsymbol{s}_i))$ is the success probability at $\boldsymbol{s}_i$, defined by a link function, and where $N(\boldsymbol{s}_i)$ represents the fixed number of trials. A logit link function specifies $p(\eta(\boldsymbol{s}_i)) = \exp(\eta(\boldsymbol{s}_i))/(1 + \exp(\eta(\boldsymbol{s}_i)))$. In some cases, the exponential family density could also include an unknown vector of nuisance parameters, say $\boldsymbol{\psi}$. If they arise, we would simply modify our hierarchical model to accommodate a prior for $\boldsymbol{\psi}$. A more general unimodal non-Gaussian likelihood, outside the exponential family class, would also fit into our framework.

3. Approximate Bayesian inference

In this section we apply approximate Bayesian inference methods to the predictive process models. Our established notation in Section 2 allows us to use INLA directly.

3.1. The Laplace approximation for predictive process model

MCMC algorithms are the current standard for inference in hierarchical Bayesian models. The generality of MCMC allows fitting very flexible models. One challenge with MCMC is the slow mixing that can occur, such that subsequent samples in the Markov chain are very dependent, and a huge number of MCMC iterations are required to explore the sampling space and to reduce the Monte Carlo error bounds sufficiently.

The Laplace approximation, see Tierney and Kadane (1986), was constructed for deterministic Bayesian inference, not for sampling based inference. The approach presented in Rue et al. (2009) is in the same spirit. Rather than sampling, analytical Gaussian approximations and numerical routines are applied. The Gaussian approximation is used for the latent effects, which are a priori Gaussian. Using notation from (11) and (12) in the previous section, we have prior $\pi(\boldsymbol{v}^* \mid \boldsymbol{\theta}) = N(\boldsymbol{v}^* \mid \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{v}^* = (\boldsymbol{w}^{*\prime}, \boldsymbol{\beta}')'$. We now let $\boldsymbol{\theta}$ denote the covariance parameters. With a Gaussian predictive process model the set of parameters is $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\phi}, \tau^2)$, while $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\phi})$ for the predictive process GLMM formulation. The LA approach exploits a recombination of the marginals and conditionals so that

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \frac{\pi(\boldsymbol{Y} \mid \boldsymbol{v}^*, \boldsymbol{\theta})\pi(\boldsymbol{v}^* \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{Y})\pi(\boldsymbol{v}^* \mid \boldsymbol{Y}, \boldsymbol{\theta})}, \\ \propto \frac{\pi(\boldsymbol{Y} \mid \boldsymbol{v}^*, \boldsymbol{\theta})\pi(\boldsymbol{v}^* \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{v}^* \mid \boldsymbol{Y}, \boldsymbol{\theta})}, \qquad (13)$$

where the numerator is defined by the model, while the left hand side and the denominator are needed for posterior inference. The full conditional in the denominator of (13) is Gaussian for a normal likelihood model (Section 2.1). Then, the posterior $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y})$ in (13) can be evaluated exactly, up to a normalizing constant. For posterior inference about these covariance hyperparameters we turn to numerical methods. This formula is identical to the marginalized model in (9). The LA approach is in this way a marginalization method using the full conditional, rather than integrating out the latent effects, see (4).

When the likelihood model is non-Gaussian, such as in the GLMM, the full conditional $\pi(\boldsymbol{v}^* \mid \boldsymbol{Y}, \boldsymbol{\theta})$ is no longer analytically available. The LA method gives

$$\hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto \left. \frac{\pi(\boldsymbol{Y} \mid \boldsymbol{v}^*, \boldsymbol{\theta})\pi(\boldsymbol{v}^* \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{v}^* \mid \boldsymbol{Y}, \boldsymbol{\theta})} \right|_{\hat{\boldsymbol{v}}^*}, \qquad (14)$$

where $\hat{\pi}(\boldsymbol{v}^* \mid \boldsymbol{Y}, \boldsymbol{\theta})$ is a Gaussian approximation of

$$\pi(\boldsymbol{v}^* \mid \boldsymbol{Y}, \boldsymbol{\theta}) \propto \pi(\boldsymbol{Y} \mid \boldsymbol{v}^*, \boldsymbol{\theta})\pi(\boldsymbol{v}^* \mid \boldsymbol{\theta}), \qquad (15)$$

constructed to match the mode $\hat{\boldsymbol{v}}^*$ and the curvature at the mode of this full conditional. The LA gives a relative error in (14), see Tierney and Kadane (1986). The Monte Carlo error is additive, possibly giving a larger relative error in the tails.

The posterior approximation $\hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$ is explored by numerical routines. For models of reasonable complexity, the dimension of $\boldsymbol{\theta}$ is small (in our examples 2 or 3), and numerical routines can efficiently find the mode, assess uncertainty bounds, and so on. An empirical Bayes solution is obtained if we plug-in the posterior mode for $\boldsymbol{\theta}$. Instead, we construct a mass function representation of the posterior density $\hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$. The numerical routine is run for a parameterization with log

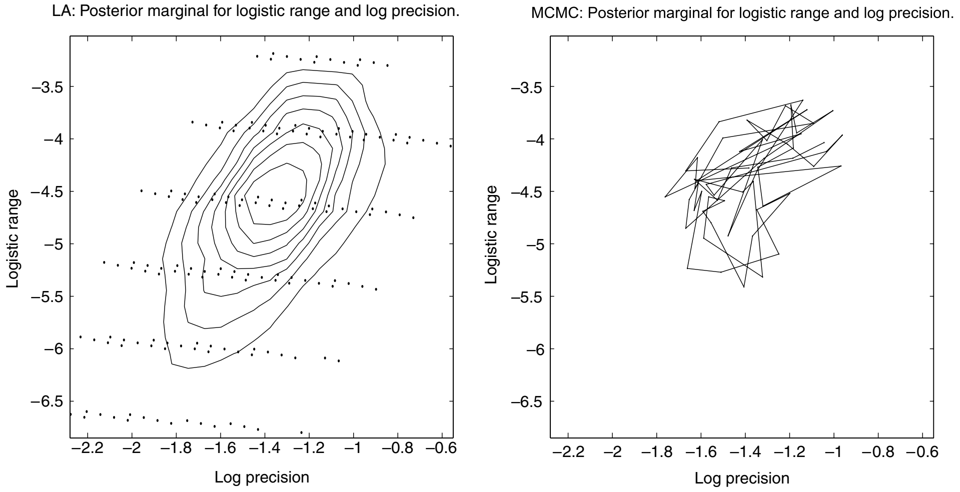

Fig. 1. Illustration of posterior inference for logistic spatial range and log precision, where the nugget parameter is marginalized out. Left: the dots are evaluation points for the numerical Laplace approach. The contours are computed based on an these points. Right: the line segments show the first 200 realizations of a MCMC run.

precision parameters and logistic spatial range. One reason for this parameterization is variance stabilization, another reason is that the surface of the approximate posterior marginal $\hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$ appears close to Gaussian and is easy to optimize. The posterior percentiles for variance or spatial decay parameters can be derived by a direct transformation. The implementation is similar to Rue et al. (2009) and goes as follows:

1. Choose a starting value $\boldsymbol{\theta}$ from prior knowledge about the scales and spatial decay.
2. Perform an optimization scheme to find the mode of $\ln \hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$.
3. Compute the Hessian of $\ln \hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$ at the mode.
4. Fill in a grid (or design) of $\boldsymbol{\theta}$ values within the region of non-negligible $\ln \hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$.
5. Evaluate and normalize $\ln \hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$ on the set of grid or design points.

In Fig. 1 (left) we show the approximate posterior marginals for log precision and logistic spatial decay parameter (integrated over the log nugget precision in this Gaussian case). The contours are constructed by rough interpolation over the evaluation points marked as dots. The pattern of evaluation points in Fig. 1 (left) is a result of the three dimensional parameter space, with one dimension (the nugget) summed out. Notice that the contours appear in an almost Gaussian/quadratic form. Altogether, the numerical optimization, Hessian computation, and gridding procedures required about $N_{la} = 200$ evaluations of the posterior. The number of evaluation points would depend on the specific goals of an application. For instance, a standard central composite design approach in the three parameter space uses only 14 evaluation points after the optimization and Hessian computation, at the cost of a coarser approximation. For comparison we display the first 200 samples of a random-walk MCMC sampler (Fig. 1, right) with acceptance probability of about 0.3. The random walk pattern is very different from the more regular pattern of the numerical scheme, and the MCMC algorithm does not span the probability space very well in the 200 iterations shown here.

We next outline the construction of the full conditional required for the Laplace approximation, in the context of a predictive process model. Every evaluation of $\hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$ entails computing this full conditional. The GLM likelihood is $\prod_i \pi(Y(\boldsymbol{s}_i) \mid \eta(\boldsymbol{s}_i))$, where the GLM parameter $\boldsymbol{\eta} = \boldsymbol{H}^*\boldsymbol{v}^* = [\mathcal{F}(\boldsymbol{\phi}), \boldsymbol{X}]\boldsymbol{v}^*$. Under Gaussian likelihood assumptions $\pi(\boldsymbol{Y} \mid \boldsymbol{\eta}) = N(\boldsymbol{H}^*\boldsymbol{v}^*, \tau^2\boldsymbol{I}_n)$, and the full conditional for $\boldsymbol{v}^*$ is

$$\begin{aligned} \pi(\boldsymbol{v}^* \mid \boldsymbol{Y}, \boldsymbol{\theta}) &\propto N(\boldsymbol{H}^*\boldsymbol{v}^*, \boldsymbol{T})N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), && (16)\\ &\propto \exp\left[-\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{H}^*\boldsymbol{v}^*)'\boldsymbol{T}^{-1}(\boldsymbol{Y} - \boldsymbol{H}^*\boldsymbol{v}^*) - \frac{1}{2}(\boldsymbol{v}^* - \boldsymbol{\mu}^*)'\boldsymbol{\Sigma}^{*-1}(\boldsymbol{v}^* - \boldsymbol{\mu}^*)\right], \\ &\propto \exp\left[-\frac{1}{2}\boldsymbol{v}^{*'}\boldsymbol{Q}\boldsymbol{v}^* + \boldsymbol{v}^{*'}\boldsymbol{b}\right], && (17) \end{aligned}$$

where $\boldsymbol{T} = \tau^2\boldsymbol{I}_n$ and the full conditional precision matrix $\boldsymbol{Q} = \boldsymbol{H}^{*'}\boldsymbol{T}^{-1}\boldsymbol{H}^* + \boldsymbol{\Sigma}^{*-1}$, i.e.

$$\boldsymbol{Q} = \begin{bmatrix} \tau^{-2}\mathcal{F}(\boldsymbol{\phi})'\mathcal{F}(\boldsymbol{\phi}) + \sigma^{-2}\boldsymbol{R}^{*-1}(\boldsymbol{\phi}) & \tau^{-2}\mathcal{F}(\boldsymbol{\phi})'\boldsymbol{X} \\ \tau^{-2}\boldsymbol{X}'\mathcal{F}(\boldsymbol{\phi}) & \tau^{-2}\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Sigma}_\beta^{-1} \end{bmatrix}. \qquad (18)$$

The canonical parameter is $\boldsymbol{b} = \boldsymbol{H}^{*'}\boldsymbol{T}^{-1}\boldsymbol{Y} + \boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}$. Thus, the full conditional is $\pi(\boldsymbol{v}^* \mid \boldsymbol{Y}, \boldsymbol{\theta}) \sim N(\boldsymbol{Q}^{-1}\boldsymbol{b}, \boldsymbol{Q}^{-1})$. The cost of matrix factorization for the predictive process model is thus $O(n^{*3})$, assuming $n^* \gg p$. The main cost of inference is the construction of $\mathcal{F}(\boldsymbol{\phi})'\mathcal{F}(\boldsymbol{\phi})$ which requires $O(nn^{*2})$.

When the likelihood model is non-Gaussian, we expand the likelihood in a quadratic form. For instance, with binomial data $\pi(Y(\boldsymbol{s}_i) \mid \eta(\boldsymbol{s}_i)) \propto \exp(Y(\boldsymbol{s}_i)\eta(\boldsymbol{s}_i) - N(\boldsymbol{s}_i)\log(1+\exp(\eta(\boldsymbol{s}_i))))$, where $N(\boldsymbol{s}_i)$ is the fixed number of trials, we Taylor expand $N(\boldsymbol{s}_i)\log(1+\exp(\eta(\boldsymbol{s}_i)))$ to second order. By expressing the result in a quadratic form of $\boldsymbol{v}^*$ we obtain

$$\log(\pi(\boldsymbol{Y} \mid \boldsymbol{\eta})) = -\frac{1}{2}\boldsymbol{v}^{*'}\boldsymbol{T}_{\text{lin}}^{-1}\boldsymbol{v}^* + \boldsymbol{v}^{*'}\boldsymbol{c}_{\text{lin}} + const, \tag{19}$$

where *const* does not depend on $\boldsymbol{v}^*$, and with

$$\boldsymbol{T}_{\text{lin}}^{-1} = \boldsymbol{H}^{*'}\boldsymbol{D}_2\boldsymbol{H}^*, \qquad \boldsymbol{c}_{\text{lin}} = \boldsymbol{H}^{*'}\boldsymbol{D}_2(\boldsymbol{Y} - \boldsymbol{d}_1 + \boldsymbol{D}_2\boldsymbol{H}^*\hat{\boldsymbol{v}}^*). \tag{20}$$

These derivative expressions are defined using component-wise multiplication ($\odot$), division ($\oslash$) and exponentiation ($\circledast$) to get

$$\begin{aligned} \boldsymbol{d}_1 &= \{\boldsymbol{N} \odot \exp(\boldsymbol{H}^*\boldsymbol{v}^*)\} \oslash \{\boldsymbol{1}_n + \exp(\boldsymbol{H}^*\boldsymbol{v}^*)\}, \\ \boldsymbol{D}_2 &= \operatorname{diag}\left(\{\boldsymbol{N} \odot \exp(\boldsymbol{H}^*\boldsymbol{v}^*)\} \oslash \{(\boldsymbol{1}_n + \exp(\boldsymbol{H}^*\boldsymbol{v}^*))^{\circledast 2}\}\right), \end{aligned} \tag{21}$$

where $\boldsymbol{1}_n$ is a $n \times 1$ vector of ones, $\boldsymbol{N} = (N(\boldsymbol{s}_1), \ldots, N(\boldsymbol{s}_n))'$, and $\exp(\cdot)$ also works componentwise. At each iteration, the linearization point is recomputed as the mode from the previous step. Five iterations are usually sufficient.

The modified predictive process models entail adding a heteroscedastic additive noise component to the process, like that described in Sections 2.1 and 2.2. In the Gaussian context the noise variances are directly added to the nugget effect, and this gives no complications for the LA. For a GLM likelihood we augment the latent state to account for the predictive process adjustment, and this gives an increase in the evaluation time of the full conditional approximations and the LA.

3.2. *The Integrated Nested Laplace Approximation for predictive process model*

The posterior marginals for regression effects or spatial effects can be computed from the Gaussian approximation of the full conditional by numerical integration over the covariance parameters. For any (spatial or regression) effect v_j^* we have

$$\hat{\pi}(v_j^* \mid \boldsymbol{Y}) = \int \hat{\pi}(v_j^* \mid \boldsymbol{Y}, \boldsymbol{\theta})\hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})d\boldsymbol{\theta}, \tag{22}$$

where $\hat{\pi}(v_j^* \mid \boldsymbol{Y}, \boldsymbol{\theta})$ is an element j of the joint approximate Gaussian. The integral is solved by numerical integration over the evaluation points for $\hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$. Numerical integration is usually superior to Monte Carlo integration in small dimensions like we have here. The full conditional for latent Gaussian variables is computed for every evaluation point of $\hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$. Thus, all entries in the integrand of (22) are readily available. This same numerical integration formula can be applied to spatial effects or a linear combination of regression parameters and spatial effects, which also has an approximate Gaussian full conditional. An empirical Bayes approach would use only one evaluation point at the posterior mode for $\boldsymbol{\theta}$.

These direct LA marginals in (22) can be improved by applying the integrated nested Laplace routine, see Rue et al. (2009) and Eidsvik et al. (2009). This INLA approach allows one particular regression effect β_j, or one spatial effect to be non-Gaussian, while all remaining latent effects remain Gaussian. The approach is based on

$$\pi(v_j^* \mid \boldsymbol{Y}, \boldsymbol{\theta}) \propto \frac{\pi(\boldsymbol{Y} \mid \boldsymbol{v}^*, \boldsymbol{\theta})\pi(\boldsymbol{v}^* \mid \boldsymbol{\theta})}{\pi(\boldsymbol{v}_{-j}^* \mid v_j^*, \boldsymbol{Y}, \boldsymbol{\theta})}, \tag{23}$$

where the latent effect v_j^* can again be a regression effect, spatial effect, or a linear combination. We will denote the resulting approximation by $\tilde{\pi}(v_j^* \mid \boldsymbol{Y}, \boldsymbol{\theta})$. Its computation uses (23) with a Gaussian approximation $\hat{\pi}(\boldsymbol{v}_{-j}^* \mid v_j^*, \boldsymbol{Y}, \boldsymbol{\theta})$ in the denominator, treating v_j^* as fixed (measured). Thus, the improved approximate marginal becomes

$$\tilde{\pi}(v_j^* \mid \boldsymbol{Y}, \boldsymbol{\theta}) \propto \left. \frac{\pi(\boldsymbol{Y} \mid \boldsymbol{v}^*, \boldsymbol{\theta})\pi(\boldsymbol{v}^* \mid \boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{v}_{-j}^* \mid v_j^*, \boldsymbol{Y}, \boldsymbol{\theta})} \right|_{\hat{\boldsymbol{v}}_{-j}^* \mid v_j^*}. \tag{24}$$

This expression is evaluated at the full conditional mode, keeping v_j^* at a fixed value. Thus, INLA uses a second round of the LA to cancel out the remaining approximate Gaussian variables $\boldsymbol{v}_{-j}^*$, for fixed v_j^*, and in this way provides a better approximation for the posterior marginals for spatial effects and regression effects, see Rue et al. (2009). The improved approximation $\tilde{\pi}(v_j^* \mid \boldsymbol{Y}, \boldsymbol{\theta})$ can be computed on a grid of v_j^* values, or fitted by a parametric density. For instance, a Gaussian approximation requires only three evaluation points v_j^* to assign a mean and covariance, and normalize.

The posterior marginal obtained by INLA becomes

$$\tilde{\pi}(v_j^* \mid \boldsymbol{Y}) = \int \tilde{\pi}(v_j^* \mid \boldsymbol{Y}, \boldsymbol{\theta})\hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})d\boldsymbol{\theta}, \tag{25}$$

which is solved by numerical integration over the same evaluation points of the covariance parameters $\boldsymbol{\theta}$. In our experience, INLA provides a shift of the LA for regression parameters $\boldsymbol{\beta}$, but very little for the spatial effects $\boldsymbol{w}$. Intuitively, the non-Gaussian data has greater effect on the regression parameters, which are valid for the entire spatial domain. The spatial effects are local variables and learn effectively from only parts of the data. Thus, the Gaussian prior model gets more dominating for spatial prediction, and the LA is more accurate.

4. Analysis and results

Five datasets are used to explore the candidate models' ability to estimate parameters of interest and predict at new locations. The first two are synthetic. The third has been used to understand the distribution of lake acidity and subsequent decline in trout abundance in Norway. These first datasets are moderate in size, and a full MCMC based data analysis is possible to perform. This allows comparison between MCMC and INLA, and between different predictive process models. The fourth and fifth are large forest inventory datasets used to produce estimates of forest land use and biomass across north-central United States. For these datasets, full process modeling and MCMC based inference is computationally prohibitive for a modern desktop workstation. The following subsections describe these datasets and accompanying modeling details.

The LA and INLA based analyses were conducted using Matlab version 7.9. Examples of this code are available at www.math.ntnu.no/~joeid/INLA_PP. The R package *spBayes* was used for MCMC based analyses. In this package the higher level R code calls C++ and Fortran that subsequently calls BLAS (www.netlib.org/blas) and LAPACK (www.netlib.org/lapack) routines for efficient matrix computations. All analyses were conducted on a Linux workstation using two Intel Nehalem-based quad-Xeon processors. The Matlab, BLAS, and LAPACK routines were threaded and therefore leveraged multiple CPUs for matrix operations. Specifically, *spBayes* was compiled to call Intel's Math Kernel Library version 10.2 BLAS and LAPACK implementations.

4.1. Synthetic data

Datasets of Gaussian and binomial outcome variables were generated. These synthetic data are composed of 750 locations selected randomly from a $[1, 100] \times [1, 100]$ square. The eight candidate models include a full geostatistical and three predictive process specifications for both MCMC and LA methods. The predictive process models are based on 64, 100, and 256 regular grid knot intensities covering the square.

Both Gaussian and binomial data were generated using a 750×3 covariates matrix $\boldsymbol{X}$, where the first column is the intercept and the values in the subsequent columns were randomly generated from a $N(0, 1)$. The regression coefficients were set to $\boldsymbol{\beta} = (0.1, 0.5, 1)'$. An exponential spatial correlation function $C(w(\boldsymbol{s}), w(\boldsymbol{s}+\boldsymbol{h})) = \sigma^2 \exp(-\phi\|\boldsymbol{h}\|)$ was used, with variance $\sigma^2 = 5$ and spatial decay parameter $\phi = 0.06$, which corresponds to an effective spatial range of $\sim$50 units. Here, effective spatial range is defined as the distance (in map units) at which the spatial correlation drops to 0.05. For the continuous outcome data the nugget variance, τ^2, was set to 1. A subset of 250 observations was selected randomly to serve as a hold-out set to assess predictive performance, while parameter estimates were based on the remaining 500 observations. All candidate models were fit using the same independent prior specification with each β following a $N(0, 10\,000)$, an $IG(2, 1)$ for the variance parameters, σ^2 and τ^2, and a broad uniform support for the spatial correlation parameter $\phi \sim U(0.03, 3)$.

Inference results of the Gaussian response model are given in Table 1. Here, 'full' refers to the MCMC or INLA method that uses all the data. When the data are Gaussian, the LA is exact, and for fixed knot configurations the small differences between MCMC and LA inference in Table 1 are caused by Monte Carlo and numerical approximation errors. These differences are sometimes visible, especially for the 2.5 and 97.5 percentiles. Considering the predictive process models, the estimates for regression parameters are captured very accurately, even with 64 knots. The distributions for covariance parameters are also quite close to the results obtained using the full dataset, but the correlation range appears a little too narrow for a small number of knots, the nugget variance is slightly underestimated, while the variance in the latent process is a little overestimated. When the knot size increases to 256, the predictive process results get closer to that of full data.

The mean square prediction error (MSPE) in Table 1 is computed over the hold-out dataset. The MSPE values for MCMC and Laplace show similar increase for predictive models with few knots. This increase (about $4.5/3.5 = 30\%$ for 64 knots) is caused by the data reduction idea that plays an intrinsic role in the predictive process formulation. The main prediction differences between full size n model and predictive process models occur at hold-out sites that are close to data locations, but far from knots. A more creative design of knot locations could reduce the MSPE in this case (see e.g. Finley et al., 2009a).

The last rows in Table 1 show the number of operations and the computing time required to deliver the inference and prediction results. The number of operations is the product of the number of evaluations and the cost of every evaluation. For the full model the main evaluation cost is matrix inversion at $O(n^3)$. For the predictive process model the main cost is $O(nn^{*2})$, which is the cost of building the required size $n^* \times n^*$ matrix. The MCMC inference was based on three MCMC chains, with unique starting values, running for $N_{mcmc} = 10\,000$ iterations. The CODA package in R (www.r-project.org) was used to diagnose convergence by monitoring mixing with the Gelman–Rubin diagnostics and autocorrelations (see, e.g., Gelman et al., 2004, Section 11.6). For the LA approach we count the number of evaluations needed to reach a certain tolerance on the numerical approximation. Altogether we use about $N_{la} = 200$ posterior evaluations for the LA approach, with a simple stepping out and gridding procedure. Table 1 shows a clear reduction in the number of operations when using the predictive process models and INLA. For instance, with 64 predictive process knots, we use 60 times less operations than with the full data. Similarly, the LA approach means a factor 50 reduction in operations. The computing times appear to give slightly less gain, both for the predictive process models and for LA. This is possibly caused by implementation overhead. By merging the two ideas of predictive process models and LA we achieve sufficiently accurate results in moderate time.

In Fig. 2 (left column) we visualize the predictive process results with LA for our parameterization with log precision (top), logistic spatial decay parameter (center), and the log nugget precision (bottom). Notice that high log precision means small

Table 1
Synthetic Gaussian dataset: summary of candidate models' parameter estimates, 50 (2.5 97.5) percentiles, hold-out set mean squared error of prediction (MSEP), the number of operations and the computation time.

			MCMC Pred. proc. knots				LA Pred. proc. knots		
	True	Full	64	100	256	Full	64	100	256
β_0	0.1	−1.11 (−2.48, 0.27)	−1.07 (−2.56, 0.51)	−1.10 (−2.39, 0.28)	−1.12 (−2.38, 0.19)	−1.11 (−2.60, 0.42)	−1.10 (−2.47, 0.37)	−1.11 (−2.53, 0.40)	−1.13 (−2.39, 0.18)
β_1	0.5	0.46 (0.33, 0.58)	0.42 (0.29, 0.56)	0.44 (0.31, 0.57)	0.45 (0.33, 0.59)	0.46 (0.33, 0.58)	0.42 (0.29, 0.56)	0.44 (0.30, 0.57)	0.45 (0.32, 0.58)
β_2	1	0.99 (0.86, 1.11)	0.98 (0.85, 1.11)	0.99 (0.86, 1.12)	0.99 (0.87, 1.11)	0.99 (0.86, 1.11)	0.98 (0.85, 1.12)	0.99 (0.86, 1.12)	0.99 (0.86, 1.11)
ϕ	0.06	0.06 (0.03, 0.10)	0.05 (0.03, 0.08)	0.06 (0.04, 0.10)	0.07 (0.03, 0.11)	0.06 (0.03, 0.10)	0.06 (0.04, 0.08)	0.06 (0.04 0.09)	0.07 (0.04, 0.10)
σ^2	5	4.07 (2.72, 6.81)	4.67 (3.05, 7.89)	4.51 (3.08, 7.16)	4.13 (2.76, 6.78)	3.90 (2.60, 6.58)	4.55 (3.11, 6.96)	4.59 (3.03, 6.86)	4.05 (2.69, 6.30)
τ^2	1	1.25 (0.90, 1.62)	0.74 (0.24, 1.39)	0.68 (0.23, 1.30)	1.02 (0.36, 1.53)	1.23 (0.86, 1.68)	0.64 (0.29, 1.29)	0.67 (0.31, 1.30)	1.03 (0.55, 1.56)
MSPE		3.54	4.55	4.27	3.87	3.35	4.67	4.40	4.17
Operations (in billions)		1250	20	50	330	25	0.4	1.0	6.6
Computing time (s)		450	101	173	420	10	3	4	8

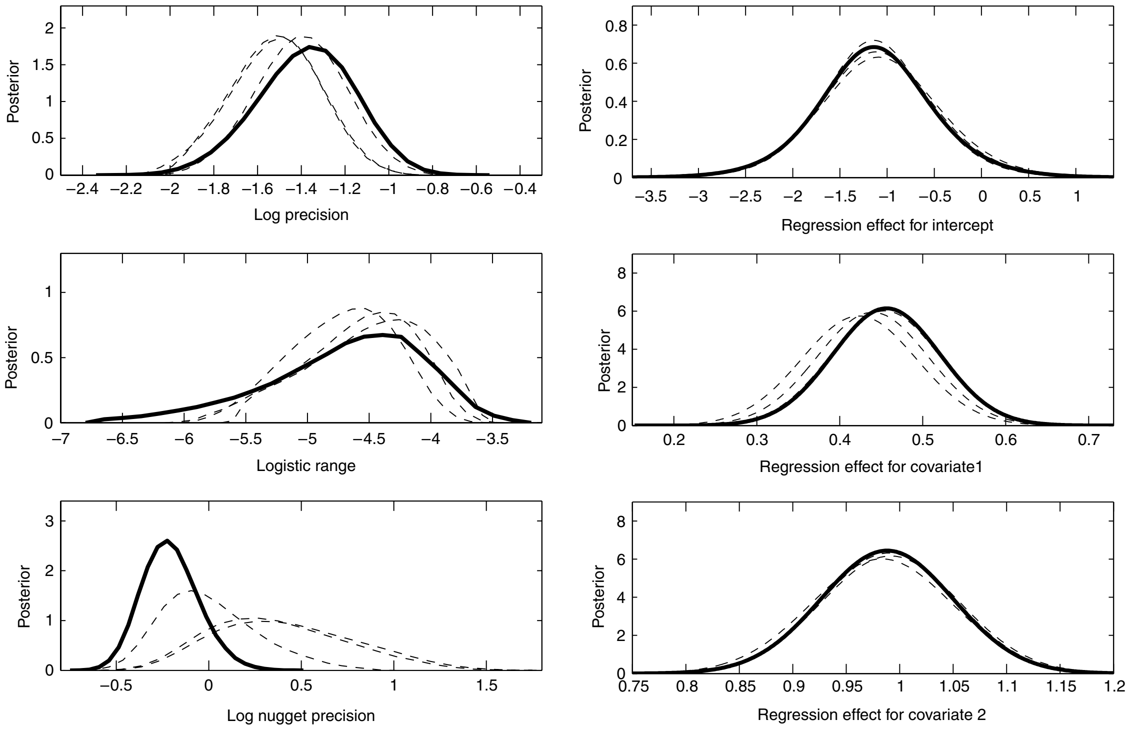

Fig. 2. Synthetic Gaussian dataset: left: LA approach for the posterior marginal of log precision (top), logistic spatial range (center), and log nugget precision (bottom). Right: LA approach for the posterior marginals of regression parameters. The dashed curves represent three different knot sizes (64, 100, 256), while the solid curve is for the full dataset ($n = 500$).

variance, so the interpretation for these parameters is opposite to that of Table 1. The displays are for the three predictive process models with 64, 100, 256 knots (dashed) and the full data (solid). The dashed lines differ a little from the solid one, but the posteriors using the predictive process get closer to the full data posterior when the knot size increases. In Fig. 2 (right column) we similarly show results for the regression effects. The predictive process models with various knot configurations and the full data provide almost the same posteriors, but with some small visible differences, especially for covariate 1 (Fig. 2, right column, center). This might be caused by quite extreme covariates at the edge of the domain, and where the knots are not so dense.

The binomial data are simulated in the same geographic locations as for the Gaussian case. In each location we draw 10 trials with the success probability at that location, using a logit link function. In Table 2 we show results of an MCMC algorithm and the INLA approach for this synthetic dataset. Two comparisons can be made in Table 2: (i) knot intensity versus full data and (ii) MCMC versus INLA. For comparing knot intensities, we see the predictive process models are quite close to the full data results, but there is some overestimation for the scale and underestimation for the spatial decay, when using few knots. This is observed both for MCMC and INLA, and was also seen for the Gaussian response model in Table 1. The effect of covariates (β_1 and β_2) are accurately estimated with the predictive process. Comparing MCMC with INLA, we see that most regression parameters and the spatial decay are very similar, while the tails of the σ^2 distribution are a little different. This could be a consequence of using a truncation rule in the numerical integration scheme for LA, but could also be caused by the Markov chain staying too long in a tail. The effect is not coupled with the particular predictive process model used. Inference for the regression parameters was done using the INLA approach. In this case we evaluated the $\tilde{\pi}(\beta_j \mid \boldsymbol{Y}, \sigma^2, \phi)$ at three evaluation points and fit a Gaussian to this improved marginal. The LA approach gives similar results as INLA, but slightly shifted up or down.

The MSPE values for the spatial effects at the hold-out set show similar tendencies as for the Gaussian dataset. The predictive process models with few knots have slightly higher MSPE. The un-marginalized models used to fit the binomial outcome data required more MCMC iterations to begin adequate mixing. The MCMC based inference in Table 2 is based on $N_{mcmc} = 25\,000$ iterations. However, each iteration is faster because of the conditional updating. The LA approach is now only over two covariance parameters (ϕ, σ^2, no nugget), and it uses fewer numerical steps than for the Gaussian data. But, on the other hand, it takes about five iterations to compute the Gaussian approximation $\hat{\pi}(\boldsymbol{v}^* \mid \boldsymbol{Y}, \sigma^2, \phi)$ for the full conditional. Thus, the number of posterior evaluations is still about $N_{la} = 200$. The operation counts in Table 2 show that the INLA solution with 64 predictive process knots uses a factor $3125/0.4 \approx 8000$ fewer operations than the MCMC sampling with full data. The gains of INLA and predictive process models are again large in terms of computation time, reflecting similar trends as for operations counts.

Please cite this article in press as: Eidsvik, J., et al., Approximate Bayesian inference for large spatial datasets using predictive process models. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.10.022

Table 2
Synthetic binomial dataset: summary of candidate models' parameter estimates, 50 (2.5 97.5) percentiles, hold-out set mean squared error of prediction (MSEP), the number of operations and computing time.

			MCMC				INLA		
			Pred. proc. knots				Pred. proc. knots		
	True	Full	64	100	256	Full	64	100	256
β_0	0.1	−0.74 (−2.49, 0.33)	−1.04 (−3.36, 0.34)	−0.99 (−3.06, 0.45)	−0.83 (−2.78, 0.39)	−0.64 (−2.13, 0.40)	−1.14 (−3.52, 0.55)	−1.11 (−3.27, 0.46)	−0.79 (−2.92, 0.77)
β_1	0.5	0.39 (0.27, 0.52)	0.42 (0.27, 0.56)	0.40 (0.26, 0.55)	0.41 (0.27, 0.55)	0.39 (0.25, 0.54)	0.39 (0.25, 0.54)	0.40 (0.26, 0.54)	0.40 (0.27, 0.54)
β_2	1	1.08 (0.95, 1.22)	1.10 (0.95, 1.25)	1.09 (0.94, 1.24)	1.09 (0.95, 1.24)	1.04 (0.87, 1.22)	1.06 (0.87, 1.22)	1.07 (0.92, 1.22)	1.08 (0.93, 1.22)
ϕ	0.05	0.07 (0.04, 0.11)	0.04 (0.03, 0.07)	0.05 (0.03, 0.08)	0.06 (0.03, 0.10)	0.07 (0.04, 0.11)	0.04 (0.03, 0.07)	0.05 (0.04, 0.08)	0.06 (0.04, 0.10)
σ^2	5	4.40 (2.93, 7.90)	5.04 (3.27, 7.55)	4.90 (3.17, 7.90)	4.65 (3.01, 8.29)	4.20 (2.78 6.96)	4.74 (3.16, 6.78)	4.60 (3.07, 6.82)	4.42 (2.91, 6.97)
MSPE		3.11	4.73	3.79	3.70	3.40	4.60	4.05	3.73
Operations (in billions)		3125	50	125	820	25	0.4	1.0	6.6
Computing time (s)		250	32	58	212	12	3	4	7

Please cite this article in press as: Eidsvik, J., et al., Approximate Bayesian inference for large spatial datasets using predictive process models. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.10.022

For both the Gaussian and binomial situations we experimented with modified vs non-modified predictive process models, and with different numerical schemes for the LA. In terms of predictions and regression effects it seems sufficient to use a non-modified predictive process and a central composite design (9 points for bivariate, 15 points for trivariate) for the $\hat{\pi}(\boldsymbol{\theta} \mid \boldsymbol{Y})$. This is a very fast alternative for inference, going beyond empirical Bayes solutions which would simply plug in the optimal $\boldsymbol{\theta}$ value. For accurate estimation of covariance parameters one benefits from using the modified predictive process and more numerical evaluation points of the LA. We also tested with different covariance ranges and correlation functions, and the results were similar to those of Tables 1 and 2.

4.2. Lake acidification

We next study a dataset originally published by Varin et al. (2005). The focus of their study was to model trout abundance in Norwegian lakes as a function of lake acidity. In a subsequent study, Hosseini et al. (2011) re-examined these data using a skewed Gaussian model and INLA for inference. The data were collected in 1986 from interviews with local fishermen. Here, we use data from the southern part of Norway. The response is 'population status' of trout for each lake $i = 1, \dots, 361$, coded as unaffected ($Y(\boldsymbol{s}_i) = 0$) or decreased/extinct ($Y(\boldsymbol{s}_i) = 1$). Lakes' northing coordinates and Acid Neutralizing Capacity (ANC) are used as covariates, along with an intercept. ANC is a measure for the overall buffering capacity against acidification for a solution.

As in the synthetic data analysis, the eight candidate models include a full geostatistical and three predictive process specifications for both MCMC and LA methods. The predictive process models are based on 54, 89, and 126 knot intensities. Table 3 shows the inference results for all candidates. For this dataset we detect almost no differences between the various predictive process models. Of course there is variability in the regression parameters and spatial decay, but considering the wide confidence bounds these differences are very small. The INLA results are similar to the MCMC, but show slight differences for the distribution of β_1 and σ^2 parameters. For the σ^2 parameter the difference in MCMC and LA seems to be driven by a heavy left tail in the MCMC results. One possible explanation is the MCMC chain stays out in the tail for too long, in the limited time of the Markov chain run. Another explanation is the truncation limits of the numerical LA approach misses this heavy tail. We constructed the INLA approximation by evaluating $\tilde{\pi}(\beta_1 \mid \boldsymbol{Y}, \sigma^2, \phi)$ at three points and fitting a Gaussian, and thus the marginal is a mixture of Gaussians. The INLA solution is almost indistinguishable from the direct Gaussian LA for the intercept and northing, while it is visibly shifted to the left for this ANC effect (β_1). Still, the INLA using a Gaussian mixture underestimates the tail a little, as extending to a non-Gaussian INLA gives a larger tail, but not quite as large as the MCMC solution. We note the posterior distribution for the spatial decay parameter almost hits the boundary for the uniform distribution for ϕ. This indicates there is limited information about the large ranges in the data, and effects of this could possibly cause some heavy tail challenges for MCMC or LA.

Fig. 3 (top) shows predictions of latent effects $\boldsymbol{x}(s)\boldsymbol{\beta} + \tilde{w}(s)$. This is displayed for the full dataset using MCMC sampling (top, left) and for the 54 knots predictive process model using the LA (top, right). For the prediction results we see little difference between these two model/inference combinations. This emphasizes that small differences in the marginals for covariance parameters or the regression effects (Table 3), translate into miniscule prediction differences in Fig. 3. There are some minor changes when going to the predictive process predictions, such as a slightly smoother result for some of the northern datapoints, that the knot configuration does not capture, but this is hard to distinguish in a map like Fig. 3 and would hardly have much effect on decision making. Similarly, Fig. 3 (bottom) shows the 95% prediction range intervals for full dataset using MCMC sampling (left) and for 54 knots using LA (right). The differences in prediction range are also small, but the full data plus MCMC results have wider ranges at some sites.

The MCMC results take a few hours to compute, while LA with 54 knots takes a few seconds. If we would like to check sensitivity to the shape of the covariance model, perform cross validation, or other such high-level inference tasks, this difference in computation time becomes important. LA with predictive process makes it possible to compute the results online on the laptop computer. We perform one such high-level task. We use cross-validation to compare the 54 knots predictive process model using LA with a non-spatial model (i.e. using only the regression part for the explanatory variables). This comparison is done over 10 randomized leave-37-out sets. For each of these sets we predict $\hat{Y}(s_{0,q})$ for every hold-out site $q = 1, \dots, 37$, using the most likely outcome in the predictive distribution as the predictor. We compare the predictions with the observed data values. Table 4 shows the results summarized by a 2 by 2 table of the total number of classifications $(\hat{Y}(s_{0,q}), Y(s_{0,q})) \in (0, 0), (0, 1), (1, 0)$, and $(1, 1)$, where a good model gets mostly the diagonal entries $(0, 0)$ and $(1, 1)$. The non-spatial model has large diagonal elements, and the explanatory variables catch much of the structure in the data. Even larger diagonal elements are achieved for the spatial model, meaning the spatial residual process adds an explanatory effect.

4.3. Forest land use

Here we analyze a large dataset where full MCMC based inference, even using the predictive process, is prohibitively expensive to run on a modern desktop workstation. The analysis is motivated by the need for spatially explicit estimates of forest area which are useful for land use change monitoring, carbon budgeting, and ecological and timber supply forecasting. The data consist of $n = 12\,629$ Forest Inventory and Analysis (FIA) plots measured in Michigan, USA, between 1999 and

Table 3
Lakes dataset: summary of candidate models' parameter estimates, 50 (2.5 97.5) percentiles.

	MCMC Pred. proc. knots				INLA Pred. proc. knots			
	Full	54	89	126	Full	54	89	126
β_0	1.72 (−1.13, 4.46)	1.58 (−1.34, 4.61)	1.73 (−1.32, 4.98)	1.76 (−1.00, 4.84)	1.59 (−1.25, 4.39)	1.50 (−1.67, 5.30)	1.64 (−1.37, 4.90)	1.62 (−1.28, 5.15)
β_1	−5.99 (−8.25, −4.39)	−6.25 (−8.95, −4.53)	−6.23 (−9.08, −4.53)	−6.20 (−8.88, −4.53)	−5.75 (−7.43, −4.26)	−5.79 (−7.36, −4.21)	−5.77 (−7.46, −4.21)	−5.77 (−7.33, −4.12)
β_2	−2.75 (−7.74, 1.57)	−2.56 (−7.75, 2.00)	−2.82 (−8.30, 1.99)	−2.80 (−8.15, 1.98)	−2.65 (−8.35, 2.14)	−2.44 (−7.77, 3.10)	−2.57 (−8.48, 2.38)	−2.57 (−7.48, 1.90)
ϕ	0.07 (0.03, 0.20)	0.07 (0.03, 0.18)	0.07 (0.03, 0.20)	0.08 (0.03, 0.22)	0.06 (0.03, 0.20)	0.06 (0.03, 0.18)	0.06 (0.03, 0.23)	0.06 (0.03, 0.24)
σ^2	2.73 (0.91, 9.02)	3.12 (0.98, 11.34)	3.20 (1.04, 12.14)	3.17 (1.03, 11.69)	2.34 (0.76, 8.15)	2.52 (0.70, 8.31)	2.49 (0.76, 8.45)	2.54 (0.72, 8.77)

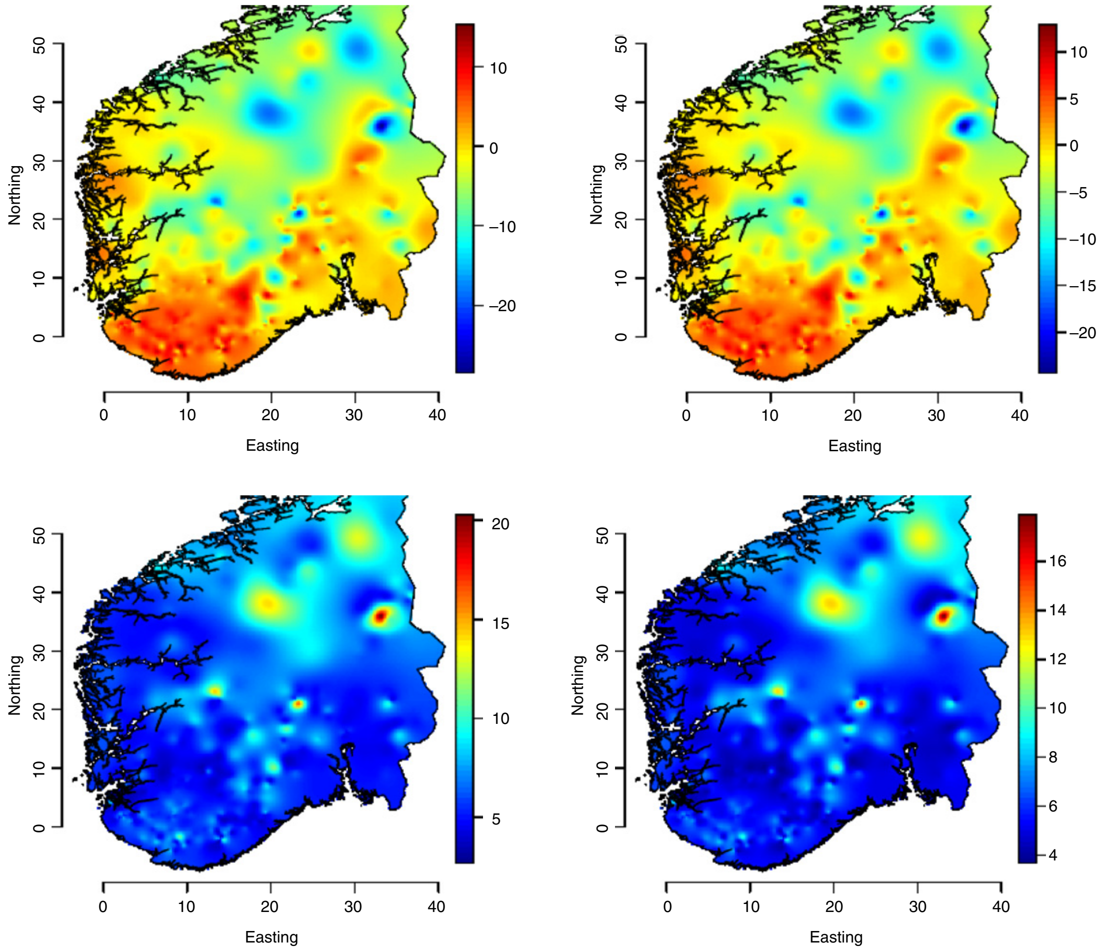

Fig. 3. Lakes dataset: prediction (top) and 95% prediction range (bottom) of the latent intensity surface. Top left: MCMC median, full dataset. Top right: LA median, 54 knot predictive process. Bottom left: MCMC range, full dataset. Bottom right: LA range, 54 knot predictive process.

Table 4
Lakes dataset: predictions and observed data for holdout sets.

	Non-spatial			Spatial Pred. proc. model, 54 knots	
	$Y(s_{0,q}) = 0$	$Y(s_{0,q}) = 1$		$Y(s_{0,q}) = 0$	$Y(s_{0,q}) = 1$
$\hat{Y}(s_{0,q}) = 0$	110	32	$\hat{Y}(s_{0,q}) = 0$	122	16
$\hat{Y}(s_{0,q}) = 1$	33	195	$\hat{Y}(s_{0,q}) = 1$	21	211

2006. The FIA program of the USDA Forest Service has established field plot centers in permanent locations using a sampling design that is assumed to produce a systematic equal-probability sample with a random spatial component, see Bechtold and Patterson (2005). Locations of plots are determined using GPS receivers. Plot locations are depicted in Fig. 4 (top left). Each plot consists of a 7.31 m radius circular area. For each plot, $i = 1, \ldots, n$, a field crew determined the response variable value as forested ($Y(\boldsymbol{s}_i) = 1$) or non-forested ($Y(\boldsymbol{s}_i) = 0$) given the FIA definition of forest land, see Bechtold and Patterson (2005). Fig. 4 (top right) is a surface interpolation of forest occupancy at the plot locations. This figure shows that northern Michigan is dominated by forest while the south is primarily not forested. A Landsat 7 ETM+ satellite image, 30 × 30 m spatial resolution, taken in mid-summer 2002 was tasseled cap transformed into its brightness, greenness, and wetness components of Kauth and Thomas (1976) and used as covariates. Finley et al. (2008) show how these components can help explain variability in the probability of forest occupancy.

Non-spatial and spatial predictive process models, using 50, 100, and 200 knot intensities, were considered in this comparison. Knot locations were chosen using a *k-means* clustering algorithm on the observed locations. Because our primary interest is in predicting forest occupancy, we compare the candidate models' using a set of 1262 holdout (or validation) plots that were selected at random from the 12 629 FIA plots. Model parameters were estimated using the remaining 11 367 observations. Table 5 summarizes the results of parameter estimation. As noted in previous studies and

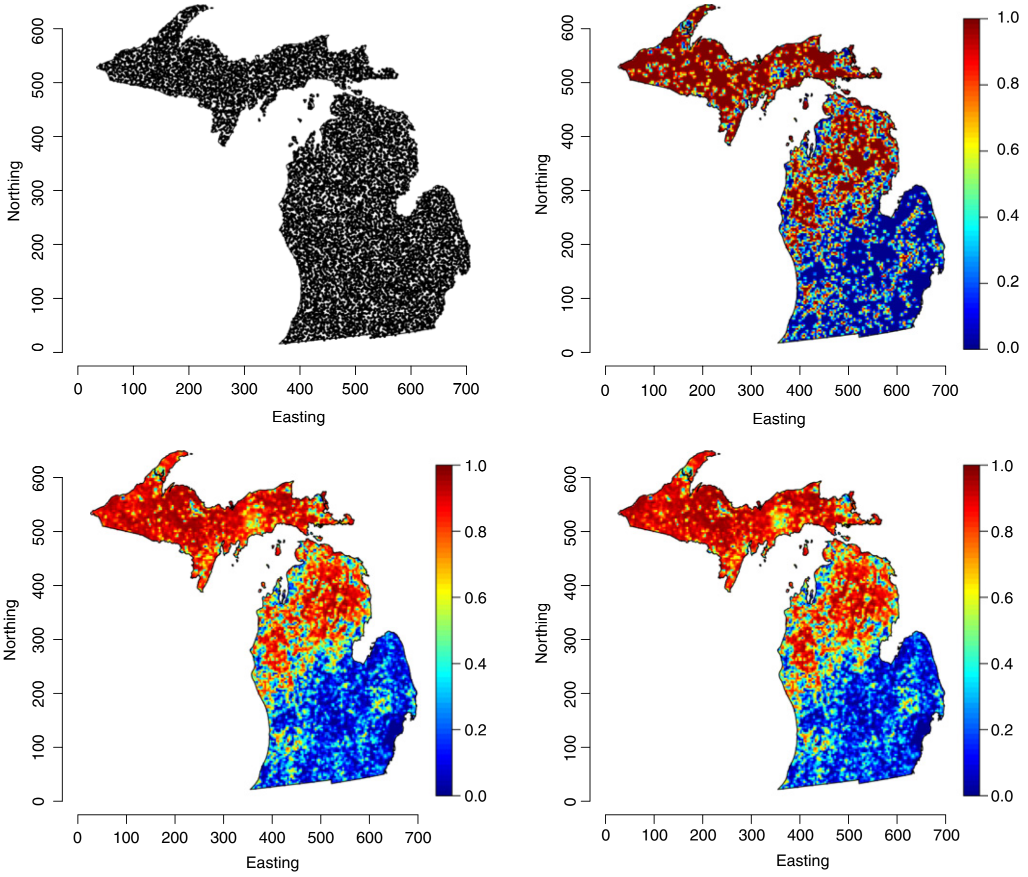

Fig. 4. Forest land use dataset: inventory data and probability of forest occupancy surfaces. Top left: forest inventory plot locations. Top right: surface of observed forest occupancy. Bottom left: median fitted probability of forest occupancy for the 50 knot model. Bottom right: median fitted probability of forest occupancy for the 200 knot model. Map units are in kilometers.

Table 5
Forest land use dataset: summary of candidate models' parameter estimates, 50 (2.5 97.5) percentiles.

	Non-spatial	INLA		
		Pred. proc. knots		
		50	100	200
β_0	4.04 (3.82, 4.28)	4.19 (2.83, 5.60)	3.86 (2.44, 5.33)	3.95 (2.23, 5.55)
β_1	−0.0044 (−0.0046, −0.0042)	−0.0049 (−0.0051, −0.0046)	−0.0049 (−0.0052, −0.0046)	−0.0049 (−0.0052, −0.0046)
β_2	0.0051 (0.0048, 0.0054)	0.0059 (0.0055, 0.0061)	0.0059 (0.0055, 0.0062)	0.0059 (0.0056, 0.0062)
β_3	−0.0027 (−0.0030, −0.0024)	−0.0033 (−0.0037, −0.0029)	−0.0034 (−0.0037, −0.003)	−0.0034 (−0.0037, −0.003)
ϕ	–	0.036 (0.033, 0.042)	0.039 (0.036, 0.045)	0.041 (0.031, 0.079)
σ^2	–	1.52 (1.40, 1.65)	1.68 (1.49, 1.89)	2.32 (1.28, 3.98)

seen here, the three remotely sensed covariates contribute significantly to explaining variability in the probability of forest occupancy and do not change substantially among the candidate models. The predictive process models produce similar estimates of the spatial decay, ϕ, and variance, σ^2, parameters. The median effective spatial ranges (in km) are 83, 76, and 73 for the 50, 100, and 200 knot model, respectively. This large spatial range is capturing the broad scale residual dependence within the forested (north) and prairie/agriculture (south) land use patterns. These results suggest we could potentially use fewer than 50 knots; however, the minimum inter-site distance among the knots of the 50 knot model is 45.6 km. Models using fewer knots, and hence greater distance between knots, could have trouble estimating the spatial decay parameter. The bottom two plots in Fig. 4 depict surface interpolation of the 50 and 200 knot models' median fitted probability of forest at the observed locations. The two models produce nearly indistinguishable surfaces, both of which capture the patterns in the non-statistical surface generated using the actual observations (top right).

Because our primary interest is in predicting forest occupancy, we compare the candidate models using the 1262 holdout sites. Extending the zero–one prediction rule from the Norwegian Lakes dataset in the previous section, we now consider several different scoring rules to evaluate the predictive performance of the candidate models. A scoring rule provides a

Table 6
Forest land use dataset: mean scoring rules based on prediction of hold-out set.

	Non-spatial	INLA		
		Pred. proc. knots		
		50	100	200
Zero-one	0.80	0.88	0.88	0.88
Quadratic	−0.29	−0.18	−0.18	−0.18
Spherical	0.84	0.90	0.90	0.90
Logarithmic	−0.45	−0.31	−0.30	−0.30

summary measure for evaluating a probabilistic prediction given the predictive distribution and the observed outcome. In our setting the scoring rule function is $SR(\boldsymbol{\pi}, i)$, where $\boldsymbol{\pi}$ is the vector of probabilities associated with each category (here $\boldsymbol{\pi}$ is of length $J = 2$, i.e., forest and non-forest) and $i = Y(\boldsymbol{s}_0)$ is the observed condition at a hold-out site. Given all the holdout sites $\{\boldsymbol{s}_{0q}\}_{q=1}^{1262}$ we can calculate summary statistics of the scores, e.g., the mean score is $\widehat{SR} = \sum_{q=1}^{1262} \frac{SR(\boldsymbol{\pi}_q, i_q)}{1262}$, where $\boldsymbol{\pi}_q = \{P(Y(\boldsymbol{s}_{0q}) = 0 \mid \boldsymbol{Y}), P(Y(\boldsymbol{s}_{0q}) = 1 \mid \boldsymbol{Y})\}$. Gneiting and Raftery (2007) offer four scoring rules for the prediction of categorical variables:

$$\text{Zero–one: } SR(\boldsymbol{\pi}, i) = \begin{cases} 1 & \text{if } \pi_i = \max\left\{\pi_1 \ldots \pi_J\right\}, \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{Quadratic: } SR(\boldsymbol{\pi}, i) = 2\pi_i - \sum_{j=1}^{J} \pi_j^2 - 1,$$

$$\text{Spherical: } SR(\boldsymbol{\pi}, i) = \frac{\pi_i}{\left(\sum_{j=1}^{J} \pi_j^2\right)^{\frac{1}{2}}},$$

$$\text{Logarithmic: } SR(\boldsymbol{\pi}, i) = \log \pi_i.$$

Following definitions in Gneiting and Raftery (2007), all the noted scoring rules are strictly *proper* but for the zero–one, which is only proper. The zero–one scoring rule uses only a portion of available information, ignoring variability in the predictive distribution and returning either a zero or one. Similarly, the logarithmic scoring rule considers only one of the probabilities in the predictive distribution. The maximum values (i.e., perfect prediction) of the different scoring rules are 1 for zero–one, 0 for quadratic, 1 for spherical, and 0 for logarithmic. The results are shown in Table 6. Here, for all rules, the spatial models offer improved predictive performance. Further, there seems to be only limited gain in predictions between the 50 and 200 knot model. Naturally, the log scoring rule penalizes the most.

4.4. *Forest biomass*

This analysis is motivated by the need for spatially explicit estimates of forest biomass that are used for contemporary global-, regional-, and local-scale decisions, including assessments of current carbon stock and flux, bio-feedstock for emerging bio-economies, and impact of deforestation. We again use the FIA dataset introduced in Section 4.3, but now consider a much larger domain that includes Michigan, Wisconsin, Minnesota, Iowa, Illinois, Indiana, and Missouri. This expanded dataset consists of $n = 60\,000$ plots where forest biomass was measured. Plot locations are depicted in Fig. 5 (left), while (middle) is a surface interpolation of biomass at the plot locations.

A non-spatial regression and two spatial predictive process models, using 64 and 256 knot intensities, were considered in this comparison. Knot locations were again chosen using a *k-means* clustering algorithm on the observed locations. We use a remotely sensed covariate, Normalized Difference Vegetation Index (NDVI), which is a measure of land surface greenness that has been shown to have a positive correlation with forest biomass (see e.g. Dong et al., 2003). For this analysis NDVI was derived from the Moderate Resolution Imaging Spectroradiometer (http://glcf.umiacs.umd.edu/data/modis) sensor at a 250 × 250 spatial resolution. To more closely approximate a Gaussian distribution, the plot-level measurements of metric tons of biomass per ha were log-transformed prior to the analysis. A subset of 10 000 plots were withheld for model validation. Model parameter estimates were then based on the remaining 50 000 observations. We assumed an exponential spatial correlation function. A preliminary analysis suggested the variance parameters followed a $IG(2, 0.75)$ and the spatial decay ϕ followed $U(0.003, 0.3)$, a priori. We used Gaussian priors with large variances for the two regression parameters (intercept term and NDVI).

The 64 and 256 knot predictive process models delivered parameter and predictive inference in 16 and 23 h, respectively. Table 7 provides parameter estimates and holdout set MSPE for the three candidate models. Similar to the previous analyses, there is very little difference across the regression parameter estimates. In all models, and consistent with previous studies, NDVI explains a significant portion of variability in biomass. Based on the MSPE, the addition of spatial random effects

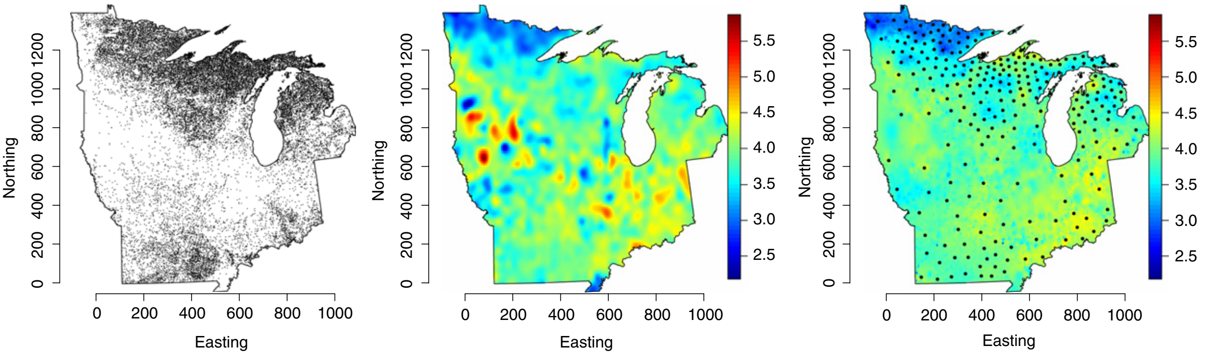

Fig. 5. Forest biomass dataset: inventory data and forest biomass. Left: forest inventory plot locations. Middle: surface of observed log forest biomass. Right: estimated median log forest biomass per ha for the 256 knot model with knot locations indicated with point symbols. Map units are in kilometers.

Table 7
Forest biomass dataset: summary of candidate models' parameter estimates, 50 (2.5 97.5) percentiles and holdout set mean squared prediction error.

	Non-spatial	INLA	
		Pred. proc. knots	
		64	256
β_0	2.40 (2.28, 2.53)	2.64 (2.12, 3.09)	2.63 (2.10, 3.15)
β_{NDVI}	6.46 (5.86, 7.10)	5.93 (4.13, 7.72)	5.86 (4.06, 7.66)
ϕ	–	0.003 (0.003, 0.004)	0.007 (0.005, 0.009)
σ^2	–	0.21 (0.17, 0.27)	0.20 (0.16, 0.26)
τ^2	1.20 (1.18, 1.21)	0.20 (0.16, 0.27)	0.21 (0.15, 0.26)
MSPE	0.81	0.67	0.65

improved predictive performance. Among the predictive process models, a larger number of knots provided a marginally better approximation to the residual spatial surface and hence greater predictive accuracy—MSPE of 0.67 versus 0.65 from the 64 and 256 knot models respectively. Although the variance parameter estimates among the spatial models are similar, the shorter spatial decay estimate from the 256 knot model, median effective spatial range $-\log(0.05)/0.007 \approx 428$ km, produces random effects with more local detail and less large-scale smoothing than the 64 knot model. Fig. 5 (right) shows a surface interpolation of the 256 knot model's median fitted log biomass at the observed location.

5. Discussion and guidelines for use

The predictive process model is built on optimal spatial prediction from a low-rank representation of the process. It introduces no extra model parameters, and, as we have shown here, goes very well with the INLA method for fast Bayesian inference. Some other knot-based methods rely on multi-resolution 'kernels'. This could give more flexibility, with the cost of difficult parameter estimation. Moreover, the resulting model is not always consistent with the Gaussian parent process. In contrast, the predictive process is the optimal 'kriging' representation of the parent process at the knots.

We limit attention to latent Gaussian models with exponential family likelihood. This gives a unimodal posterior. Otherwise, the INLA approach, using a Gaussian approximation for the latent process, would focus on the local modes. Within this model class the predictive process is used as a representation for the latent Gaussian process.

In our simulations we tried to pin down sources of error by using the same INLA approach for different size predictive process models, and using various numerical integration (INLA) approaches for one predictive process model. A main contribution of our work is that we found no interaction effects that make the combined approximations break down. The error of INLA is relative and MCMC sampling is additive, but the estimation and prediction results are similar across many predictive process models. For the other situation, the predictive process model gives similar results for a range of different numerical integration techniques used in the INLA approach. The hardest cases would occur when there are several spatial decay parameters. First, the predictive process model uses no adjustment for the decay parameters in its approximation. Moreover, the numerical integration scheme used by INLA requires a moderate number of covariance parameters. Thus, with a very complicated covariance model, the numerical integration would have to be coarse, and the combined effect of using the predictive process model and INLA might produce compromised results.

Recall that the predictive process model is inherently non-stationary even when the parent process is stationary. This warrants that its parameters are interpreted within the context of a nonstationary model. It is, therefore, not surprising that some of the parameter estimates are a little different from what would have been obtained by fitting the parent model. However, the inference and predictions rarely reveals substantive changes.

Here are some guidelines for using predictive process models and INLA in large spatial models:

- The standard non-modified predictive process model performs adequately for prediction and the estimation of regression effects. For accurate assessment of the posterior of covariance parameters we recommend using more knots and applying the modified process.
- The yardstick of predictive process performance is to try different numbers of knots, within a computationally feasible range, and to check knots sensitivity for obtaining reliable results. The following knots designs are useful: regular placement of knots covering the domain, or a k-means clustering placement focusing knots where there is more data.
- The standard LA performs adequately for prediction of spatial effects. For the posterior of regression parameters we recommend using INLA, which tends to give a slight shift to the posterior density for these fixed effects.
- Numerical assessment by a few central composite design points gives sufficient results for regression and spatial effects. For accurate assessment of the posterior for covariance parameters we recommend a denser and wider resolution of the covariance parameter space. If the main interest is some function of the covariance parameters, then one could reparameterize the model such that this one parameter becomes a key feature in the numerical routine.

6. Conclusions

The main contribution of this paper is combining computational ideas for modeling and inference of spatial data. The paper synthesizes the predictive process models and approximate Bayesian inference using INLA. This provides fast and reliable analysis of large spatial datasets. The predictive process models entail a selection of knot locations that covers the domain of interest. As the number of knots is much smaller than the number of data sites, matrix computations are feasible, even in very high dimensions. The Laplace approximation, combined with numerical routines, provides much faster inference than MCMC algorithms for spatial geostatistical models. The predictive process models and the INLA approach go well together because the predictive process is a dimension reduction technique which introduces no extra covariance model parameters. No added theoretical or computational costs are introduced by combining the two approaches, and we noticed no interaction effects that could cause the combined approximations to break down.

Further work might include predictive process modeling and INLA for multivariate spatial data, space–time applications, or to situations where the regression parameters change in space–time, such as the spatially varying coefficients model. The computational problems are further aggravated in these situations, and combining predictive process models and INLA appears very appealing.

Acknowledgments

The authors thank the Statistical and Applied Mathematical Sciences Institute (SAMSI) for hospitality during the 2009–2010 program on Space–Time Analysis for Environmental Mapping, Epidemiology and Climate Change. The work of the second author was supported by the USDA Forest Inventory and Analysis National Program and Forest Health Technology Enterprise Team. The second and third authors were supported by the National Science Foundation (NSF) grant NSF-DMS-1106609 and National Institutes of Health (NIH) grant NIH/NIGMS 1-RC1-GM092400-01. The first and fourth author acknowledge project support from the Research Council of Norway.

References

Ainsworth, L.M., Dean, C.B., 2006. Approximate inference for disease mapping. Computational Statistics and Data Analysis 50, 2552–2570.
Banerjee, S., Carlin, B.K., Gelfand, A.E., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall.
Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial datasets. Journal of the Royal Statistical Society, Series B 70, 825–848.
Banerjee, S., Finley, A.O., Waldmann, P., Ericsson, T., 2010. Hierarchical spatial process models for multiple traits in large genetic trials. Journal of the American Statistical Association 105, 506–521.
Bechtold, W.A., Patterson, P.L., 2005. The enhanced forest inventory and analysis program—national sampling design and estimation procedures. In: General Technical Report SRS-80, Asheville, NC, USDA Forest Services, Southern Research Station, 85.
Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.
Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of American Statistical Association 90, 1313–1321.
Crainiceaniu, C.M., Diggle, P.J., Rowlingson, B., 2008. Bivariate binomial spatial modeling of Loa Loa prevalence in Tropical Africa. Journal of the American Statistical Association 103, 21–37.
Cressie, N., 1993. Statistics for Spatial Data. Wiley.
Cressie, N., Johannesson, G., 2008. Fixed rank kriging for large spatial datasets. Journal of the Royal Statistical Society, Series B 70, 209–226.
Diggle, P.J., Tawn, J.A., Moyeed, R.A., 1998. Model-based geostatistics. Journal of the Royal Statistical Society, Series C 47, 299–350.
Diggle, P.J., Lophaven, S., 2006. Bayesian geostatistical design. Scandinavian Journal of Statistics 33, 53–64.
Dong, J., Kaufmann, R.K., Myneni, R.B., Tucker, C.J., Kauppi, P.E., Liski, J., Buermann, W., Alexeyev, V., Hughes, M.K., 2003. Remote sensing estimates of boreal and temperate forest woody biomass: carbon pools, sources, and sinks. Remote Sensing of Environment 84, 393–410.
Du, J., Zhang, H., Mandrekarm, V.S., 2009. Fixed-domain asymptotic properties of tapered maximum likelihood estimators. The Annals of Statistics 37, 3330–3361.
Eidsvik, J., Martino, S., Rue, H., 2009. Approximate Bayesian inference for spatial generalized linear mixed models. Scandinavian Journal of Statistics 36, 1–22.
Evangelou, E., Zhu, Z., Smith, R.L., 2011. Estimation and prediction for spatial generalized linear mixed models using high order Laplace approximation. Journal of Statistical Planning and Inference 141, 3564–3577.

Please cite this article in press as: Eidsvik, J., et al., Approximate Bayesian inference for large spatial datasets using predictive process models. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.10.022

Finley, A.O., Banerjee, S., Ek, A.R., McRoberts, R.E., 2008. Bayesian multivariate process modeling for prediction of forest attributes. Journal of Agricultural, Biological and Environmental Statistics 13, 60–83.

Finley, A.O., Sang, H., Banerjee, S., Gelfand, A.E., 2009a. Improving the performance of predictive process modeling for large datasets. Computational Statistics and Data Analysis 53, 2873–2884.

Finley, A.O., Banerjee, S., McRoberts, R.E., 2009b. Hierarchical spatial models for predicting tree species assemblages across large domains. Annals of Applied Statistics 3, 1052–1079.

Fuentes, M., 2007. Approximate likelihood for large irregularly spaced spatial data. Journal of the American Statistical Association 102, 321–331.

Furrer, R., Genton, M.G., Nychka, D., 2006. Covariance tapering for interpolation of large spatial datasets. Journal of Computational and Graphical Statistics 15, 502–523.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis. Chapman & Hall.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction and estimation. Journal of the American Statistical Association 102, 359–378.

Higdon, D., 2002. Space and space–time modeling using process convolutions. In: Anderson, C., Barnett, V., Chatwin, P.C., El-Shaarawi, A.H. (Eds.), Quantitative Methods for Current Environmental Issues. Springer-Verlag, pp. 37–56.

Hosseini, F., Eidsvik, J., Mohammadzadeh, M., 2011. Approximate Bayesian inference in Spatial GLMMs with skew normal latent variables. Computational Statistics and Data Analysis 55, 1791–1806.

Hsiao, C.K., Huang, S.Y., Chang, C.W., 2004. Bayesian marginal inference via candidate's formula. Statistics and Computing 14, 59–66.

Kammann, E.E., Wand, M.P., 2003. Geoadditive models. Applied Statistics 52, 1–18.

Kaufman, C.G., Schervish, M.J., Nychka, D.W., 2008. Covariance tapering for likelihood-based estimation in large spatial datasets. Journal of the American Statistical Association 103, 1545–1555.

Kauth, R.J., Thomas, G.S., 1976. The tasseled cap: a graphic description of the spectral–temporal development of agricultural crops as seen by Landsat. In: Proceedings of the Symposium on Machine Processing of Remotely Sensed Data. Purdue University, West Lafayette, IN. pp. 41–51.

Latimer, A.M., Banerjee, S., Sang, H., Mosher Jr., E., Silander, J.A., 2009. Hierarchical models for spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. Ecology Letters 12, 144–154.

Lewis, S.M., Raftery, A.E., 1997. Estimating Bayesian factors via posterior simulation with the Laplace–Metropolis estimator. Journal of the American Statistical Association 92, 648–655.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R., Klein, B., 2000. Smoothing spline ANOVA models for large data sets with Bernoulli observations and randomized GACV. Annals of Statistics 28, 1570–1600.

Nychka, D.W., Saltzman, N., 1998. Design of air quality monitoring networks. In: Case Studies in Environmental Statistics. Springer, pp. 51–76.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields, Theory and Applications. Chapman & Hall.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models using Integrated Nested Laplace Approximations. Journal of the Royal Statistical Society, Series B 71, 319–392.

Sang, H., Huang, J.Z., 2012. A full scale approximation of covariance functions for large spatial data sets. Journal of the Royal Statistical Society, Series B, 74, forthcoming (doi:10.1111/j.1467-9868.2011.01007.x).

Schabenberger, O., Gotway, C.A., 2004. Statistical Methods for Spatial Data Analysis. Chapman & Hall.

Stein, M.L., 1999. Interpolation of Spatial Data: Some Theory of kriging. Springer.

Stein, M.L., Chi, Z., Welty, L.J., 2004. Approximating likelihoods for large spatial datasets. Journal of the Royal Statistical Society, Series B 66, 275–296.

Stein, M.L., 2007. Spatial variation of total column ozone on a global scale. The Annals of Applied Statistics 1, 191–200.

Stein, M.L., 2008. A modeling approach for large spatial datasets. Journal of the Korean Statistical Society 37, 3–10.

Tierney, L., Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 81, 82–86.

Varin, C., Høst, G., Skare, Ø., 2005. Pairwise likelihood inference in spatial generalized linear mixed models. Computational Statistics and Data Analysis 49, 1173–1191.

Vecchia, A.V., 1988. Estimation and model identification for continuous spatial processes. Journal of the Royal Statistical Society, Series B 50, 297–312.

Xia, G., Miranda, M., Gelfand, A.E., 2006. Approximately optimal spatial design approaches for environmental health data. Environmetrics 17, 363–385.