The “*big n problem*”: a review and a simulation study

Giovanna Jona Lasinio and Gianluca Mastrantonio

Department of Statistical Sciences, Sapienza University of Rome,
P.le Aldo Moro 5, 00185 Rome

Alessio Pollice

Department of Statistical Sciences "Carlo Cecchi", University of Bari "Aldo Moro"

Abstract

When a large amount of spatial data is available computational and modeling challenges arise and they are often labeled as “*big n problem*”. In this work we present a brief review of the literature. Then we focus on two approaches, respectively based on Stochastic Partial Differential Equations and INLA, and on the tapering of the spatial covariance matrix. The fitting and predictive abilities of using the two methods in conjunction with Kriging interpolation are compared in a simulation study.

Keywords SPDE, INLA, tapering, Large Spatial Data Sets, Spatial Statistics

1 Introduction

In geophysical and environmental sciences massive amounts of data are increasingly available at a large number of spatial locations and Kriging-based spatial interpolation with parametric covariance functions is often required. For example temperature surface estimation in the context of global warming studies implies the consideration of spatial fields of huge size. In Climatology as well as in Ecology and other fields, linear spatial interpolation can successfully address spatial misalignment of available data, a serious issue that arises quite often. For instance when studying commercial fish populations, satellite data of marine physical parameters (temperature, salinity etc.) are often easily available with a spatial resolution that does not match the one of fishery data. Notwithstanding the all-time high computational power, calculations can still be an issue in these cases: the exact computation of the likelihood of a Gaussian random field observed at n spatial locations or the solution of the Kriging equations require $O(n^3)$ operations (Cressie and Johannesson (2008); Stein *et al.* (2004)). Evidence of an even more serious computational burden arises when multiple observations are available at every spatial location, as in spatiotemporal problems, or when computational expensive techniques are needed, as Markov chain Monte Carlo (McMC) methods for hierarchical model estimation and prediction in the Bayesian framework. In the Gaussian setting, the main issue is the computation of quadratic forms involving the inversion of covariance matrices that has to be repeated over and over in McMC based estimation. As clearly illustrated in Banerjee *et al.* (2004) (appendix A.5), even if we do not need the inverse of the covariance matrix in order to compute the quadratic form $\mathbf{p}'\mathbf{A}^{-1}\mathbf{p}$, the burden of large n is still there: though we can obtain $\mathbf{A}^{1/2}$ and solve $\mathbf{p} = \mathbf{A}^{1/2}\mathbf{v}$ for $\mathbf{v}$, then $\mathbf{v}'\mathbf{v} = \mathbf{p}'\mathbf{A}^{-1}\mathbf{p}$, still computation with associated $n \times n$ matrices can be unstable and repeated computation can be extremely slow.

In recent years computational issues giving rise to the so called "*big n problem*" were approached by the proposal of a number of statistical methods to improve the computational efficiency and make the analysis of large spatial datasets feasible (Banerjee *et al.*, 2004). Sun *et al.* (2012) pointed out that there is a growing need of works that compare different methods for handling large spatial datasets with simulation studies and based on real data examples. In this work we first review the main approaches in a unified framework in section 2, highlighting pro's and con's and giving pointers to the relevant literature. Sections 3 and 4 are focused on two approaches to Kriging-based interpolation with parametric spatial covariance functions. In section 3 we summarize the key findings proposed by Lindgren *et al.* (2011) and briefly illustrate the SPDE/INLA approach. In section 4 we describe the tapering approach and some relevant results. Section 5 is devoted to two simulation experiments comparing the fitting and predictive ability of the two approaches, and section 6 contains some concluding remarks.

2 Review of some relevant literature on the "*big n problem*"

In the last two decades the "*big n problem*" was addressed by several approaches including subsampling, spectral methods, likelihood and covariance function approximations, process convolutions, Markov random field representations, solutions to a class of stochastic partial differential equations (SPDE's), projections in a subspace of reduced dimension and non-parametric methods. None of the proposed approaches fits all the possible situations and problems can arise even when the assumptions underling a specific methodology are fully satisfied.

The use of an appropriate subsample of the observed locations can be considered as a first attempt to approach the "*big n problem*". Though ignoring some of the available data might look unattractive, Banerjee *et al.* (2004) argue that a clever subset of the sampling locations may allow better inferences than the use of all data. In the complete data set the possible presence of locations very close to one another with strong spatial dependence leads to a nearly singular covariance matrix for a purely spatial model (without nugget effect). Then working with a subsample can ease estimation of model parameters and reduce the general impact of numerical stability problems. Ad hoc methods of subsetting the data were formalized as in the moving window approach by Haas (1995), although it appeared that the local covariance functions fitted within the windows yielded incompatible covariances at larger spatial lags.

Large spatial data sets not only rise computational issues, but are also challenging from a modeling point of view. Such datasets are often defined on a large spatial domain and the spatial processes of interest typically exhibit a non-stationary behavior. An approach to deal with the "*big n problem*" in this case, limited by the requirement of multiple observations at every site, is given in Sampson and Guttorp (1992). Their deformation method for non-stationary spatial data maps the original process into a new space where the covariance is a function of the distances between sites (the process is stationary). With large spatial domains, the transition from a non-stationary to a stationary isotropic spatial covariance usually increases the number of elements of the covariance matrix small enough to be considered negligible and set to zero. Then if the number of zeros is large enough (covariance with a small practical range), sparse matrix algorithms can be used to enhance the computational speed.

An alternative class of methods is based on the representation of the n-dimensional spatial process by a lower-dimensional latent process. One of the first proposals along these lines combines the Kalman filter and Kriging in a spatio-temporal model (Mardia *et al.*, 1998), where the latent spatial process at time t is obtained by the low-dimensional state vector of the Kalman filter multiplied by a vector of parameters. More recently, Cressie and Johannesson (2008) used a low-rank representation to obtain a dimensional reduction of the covariance matrix in a Kriging framework. Their fixed rank Kriging model maps the original process into a new lower-dimensional process through the definition of a specific class of covariance functions C defined as the product of basis functions and an $m \times m$ positive definite matrix $\mathbf{D}$, with $m \ll n$, i.e. $C(\mathbf{s}_i, \mathbf{s}_j) = \mathbf{S}(\mathbf{s}_i)'\mathbf{D}\mathbf{S}(\mathbf{s}_j)$, where $\mathbf{S}(\cdot)$ is a $m \times 1$ vector of basis functions and $\mathbf{s}_i \in \mathbb{R}^2$ with $i, j = 1, \dots, n$. Matrix $\mathbf{D}$ is estimated by a method based on minimizing a weighted Frobenius norm and the quality of the fixed rank Kriging approximation depends on m and on the choice of the basis functions. The authors propose to choose among spline, wavelet and radial basis functions, adding as a general recommendation the use of multi-resolution basis functions.

The idea of using a low-dimensional representation of the process to decrease the computational burden is also the baseline of the kernel convolution methods that express a stationary Gaussian process $z(\cdot)$ as a function of a two dimensional Brownian motion b and a bivariate kernel k. While kernel convolution was originally used to represent stationary processes, Higdon *et al.* (1998) proposed a kernel convolution model with kernel depending on the spatial location and the following discrete representation:

$$z(\mathbf{s}_i) = \int_{R^2} k_i(\mathbf{s}_i - \mathbf{t})\, b(\mathbf{t})\, d\mathbf{t} = \sum_{j=1}^{m} k_i(\mathbf{s}_i - \mathbf{t}_j)\, b(\mathbf{t}_j).$$

A computational gain is obtained when the dimensional reduction is such that $m \ll n$. Inferences with the kernel convolution methods are influenced by the set $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_m)$, though literature does not provide any advice for the choice.

In the same spirit the knot model proposed by Banerjee *et al.* (2008) achieves the dimensional reduction with the choice of a specified set of $m \ll n$ locations (knots). The resulting class of models essentially falls under the generalized linear mixed model framework and includes kernel convolution methods as a special case. The original process is projected onto a subspace generated by realizations of the process at the knots. Approximation properties are highly sensitive to the number and location of the knots: the value m has to assure feasible computational times, while knot location is addressed using a regular grid over the domain or alternatively by placing more knots where there are more observations. Further improvements on the knot model are given in Finley *et al.* (2009).

For stationary Gaussian data on a regular lattice, Whittle (1954) proposed a method to estimate the spectrum of the spatial process and to approximate the log-likelihood using the spectral density and the periodogram. The fast Fourier transform is used to obtain the approximation, with order $O(n \log_2 n)$ computational complexity. As the regular lattice is a strong restriction, there were proposals to use the Whittle approximation with irregularly spaced spatial data. Fuentes (2007) and Matsuda and Yajima (2009) proposed parametric and non-parametric methods to estimate the process likelihood in the spectral domain; the major difference between the two approaches is that the first assumes the spatial locations to be fixed and possibly missing at random, while the second considers them realizations of a random variable with known distribution.

Another possible way to deal with the "*big n problem*" is to approximate a Gaussian field (GF) with a Gaussian Markov random field (GMRF) for which more computationally feasible algorithms for sparse matrices can be used. Initially GMRF approximations were heavily criticized for several reasons: they don't allow a direct modeling of spatial associations, precluding the consideration of specific correlation behaviors; we cannot write down the distribution at a selected location, but rather determine a conditional distribution given a prespecified set of locations; introduction of nonspatial error is confusing (see Banerjee *et al.*, 2004, and references therein). The above criticisms were recently overcome by Lindgren *et al.* (2011) who introduced the GMRF approximation of a GF with Matérn covariance function as a solution of a Stochastic Partial Differential Equation (SPDE). The SPDE approach allows a rigorous evaluation of the approximation error and is applicable to regularly and irregularly spaced spatial data, even on a curved surface. Essential to the development of this approach are the works of Rue and Tjelmeland (2002) and Rue and Held (2005) on the integrated nested Laplace approximation (INLA), whose main output is the INLA software (Rue *et al.*, 2009). INLA assumes the observed response variable to be distributed as a member of the exponential family with expectation linked to a structured additive predictor including a latent GF, possibly a GMRF approximation of a GF (see section 3 for details and Lindgren *et al.* (2011)). The use of the Laplace approximation ensures that the posterior distribution of the parameter vector has error rate of order $O(n^{-1})$ (for more information about the approximation error in INLA see section 4 in Rue *et al.* (2009)). INLA allows to do Bayesian computations in a fraction of the time needed for McMC, but has some major limitations: the dimension of the parameter vector has to be smaller than six to obtain feasible computational times, it is not capable to handle multimodal distributions of the parameter vector and the approximation is poor for highly non-Gaussian likelihoods. Some recent proposals aiming at minimizing the limitations of the INLA implementation are given in Ji *et al.* (2011) where the IS-LA, MH-LA an MLA approaches are introduced, that combine Importance Sampling (Rubinstein, 1981), Metropolis Hastings and a Marginalization based approach with the Laplace approximation respectively.

Finally sparse spatial covariance matrices can be obtained by truncating the covariance function to zero for distances farther than a given value. This is the basic idea behind the tapering method which multiplies the covariance function of the spatial process by a tapering function, in order to obtain an equivalent process with a sparse covariance matrix (see section 4 for details and Furrer *et al.* (2006)). The literature on the tapering provides two different kinds of approximations: one that has simpler asymptotic results but biased maximum likelihood estimates, another one that leads to unbiased maximum likelihood estimates (Kaufman *et al.*, 2008). Asymptotic results are provided for both approximations in the increasing domain, i.e. letting the sampling region increase without limitation together with the sample size, and in the fixed domain, where the domain is fixed and the observations become more and more dense (see for example Kaufman *et al.* (2008) or Shaby and Ruppert (2011)). A Matérn covariance is generally required to obtain a process with asymptotically valid covariance function in both the increasing and fixed domain.

In the next two sections we give some more insights on the SPDE/INLA and tapering approaches. We focus on these two methods as they both involve continuous spatial data and the assumption of Gaussian random fields, and are thus comparable in a Kriging-based estimation framework.

3 SPDE/INLA

The SPDE method was proposed by Lindgren *et al.* (2011) as an explicit and computationally efficient Markov representation of a Gaussian process $\mathbf{z} = z(s_1, \dots, s_n)$ with Matérn covariance function given by:

$$C(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa h)^{\nu} K_{\nu}(\kappa h) \tag{1}$$

where h is a generic distance between two sites, σ^2 is the process variance, K_ν is the Bessel function of order ν, $\kappa = \sqrt{8\nu}/\phi$ is a scale parameter, $\Gamma(\cdot)$ is the gamma function and $1/\phi$ is the variance decay parameter. Let Δ be the Laplacian operator in a r-dimensional space, we consider the following linear fractional SPDE

$$\left(\kappa^2 - \Delta\right)^{\frac{\alpha}{2}} z(\mathbf{s}_i) = qW(\mathbf{s}_i), \;\; \mathbf{s}_i \in \mathbb{R}^r, \quad \kappa > 0, \quad \alpha > 0, \tag{2}$$

where W is a spatial Gaussian white noise process with unit variance, $\alpha = \nu + \frac{r}{2}$, $\left(\kappa^2 - \Delta\right)^{\frac{\alpha}{2}}$ is a pseudo differential operator and

$$q^2 = \frac{\sigma^2 \Gamma(\alpha) (4\pi)^{r/2} \kappa^{2\nu}}{\Gamma(\nu)} \tag{3}$$

Whittle (1954, 1963) found that the only stationary solution of (2) has Matérn Covariance function and Lindgren *et al.* (2011) call it a Matérn field. Let $\langle f, g\rangle = \int f(\mathbf{s}_i) g(\mathbf{s}_i) d\mathbf{s}_i$ be the inner product, where the integral is over the region of interest, then the solution of (2) found requiring that

$$\left\{\left\langle \varphi_l, \left(\kappa^2 - \Delta\right)^{\alpha/2} z \right\rangle\right\}_{l=1,\dots,N} \overset{D}{=} \left\{\langle \varphi_l, qW\rangle\right\}_{l=1,\dots,N}, \tag{4}$$

for an appropriate finite set of test functions $\{\varphi_l\}$, $l = 1, \dots, N$, is called *stochastic weak solution* of the SPDE. The weak solution of (2) can be obtained by the *finite elements method* (FEM) (Brenner and Scott, 2007) and is based on the triangulation of the spatial domain into a set of non-intersecting triangles. As the quality of the solution depends on the triangulation properties, the Delaunay triangulation, which ensures smooth transitions between small and large triangles, is adopted (Lindgren *et al.*, 2011, and references therein).

The basis functions $\{\psi_l\}$, $l = 1,\dots,N$, used in the finite elements representation of the weak solution, are defined as piecewise linear basis functions:

$$\psi_i(\mathbf{t}_j) = \begin{cases} 1, \ i = j \\ 0, \ i \neq j \end{cases} \quad i, j = 1,\dots,N.$$

where $\mathbf{t}_j$ is a vertex in the triangulation. A general finite elements solution of (2) requires the calculation of matrices $\mathbf{C} = [\langle \psi_i, \psi_j \rangle]_{i,j=1,\dots,N}$ and $\mathbf{G} = [\langle \nabla\psi_i, \nabla\psi_j \rangle]_{i,j=1,\dots,N}$. While $\mathbf{G}$ is a sparse matrix, $\mathbf{C}$ is dense and Lindgren *et al.* (2011) propose to substitute it with the sparse diagonal matrix $\tilde{\mathbf{C}}$ with non-zero elements $\tilde{\mathbf{C}}_{i,i} = [\langle \psi_i, 1 \rangle]_{i=1,\dots,N}$. Then the finite elements representation proposed by Lindgren *et al.* (2011) is

$$\tilde{z}(\mathbf{t}_j) = q \sum_{l=1}^{N} \psi_l(\mathbf{t}_j) w_l, \tag{5}$$

where $\{w_l\}, l = 1,\dots,N$ are zero-mean Gaussian weights with precision matrix $\mathbf{Q}_\alpha$. The choice of the test functions $\{\varphi_l\}_{l=1,\dots,N}$ in relation to the basis functions governs the approximation properties of the resulting model representation.
With $\alpha = 1$ and $\varphi_l = \left(\kappa^2 - \Delta\right)^{\alpha/2} \psi_l$ (least squares solution)

$$\mathbf{Q}_1 = q^{-2}(\kappa^2\tilde{\mathbf{C}} + \mathbf{G})$$

With $\alpha = 2$ and $\varphi_l = \psi_l$ (Galerkin solution)

$$\mathbf{Q}_2 = q^{-2}(\kappa^2\tilde{\mathbf{C}} + \mathbf{G})\tilde{\mathbf{C}}^{-1}(\kappa^2\tilde{\mathbf{C}} + \mathbf{G})$$

For $\alpha > 2$ the precision matrix $\mathbf{Q}_\alpha$ is found with a recursive formulation based on Q_2 and Q_1 for α even and odd, respectively:

$$\mathbf{Q}_\alpha = q^{-2}(\kappa^2\tilde{\mathbf{C}} + \mathbf{G})\tilde{\mathbf{C}}^{-1}\mathbf{Q}_{\alpha-2}\tilde{\mathbf{C}}^{-1}(\kappa^2\tilde{\mathbf{C}} + \mathbf{G}).$$

As matrices $\tilde{\mathbf{C}}$, $\mathbf{G}$ and $\tilde{\mathbf{C}}^{-1}$ involved in the calculation of $\mathbf{Q}_\alpha$ are sparse, $\mathbf{Q}_\alpha$ is sparse as well and then $\{w_l\}$, $l = 1,\dots,N$, and consequently $\tilde{\mathbf{z}}$, are GMRF's. A link is then obtained between the parameters of the Matérn GF and the elements of the GMRF, with computational cost $O(N)$ for every triangulation (Lindgren *et al.*, 2011). The substitution of $\mathbf{C}$ with $\tilde{\mathbf{C}}$ is necessary because $\mathbf{C}^{-1}$ is not a sparse matrix and then the corresponding $\tilde{\mathbf{z}}$ is not a GMRF. Lindgren *et al.* (2011) show that the difference in convergence rates for the exact FEM representation and the Markov approximation is negligible.
Extensions of the SPDE approximation for a non-stationary process can be found considering the parameters κ^2 and q constant within the support of the basis function but varying slowly in space (Lindgren *et al.*, 2011).
The SPDE approximation of a GF by a GMRF is currently implemented using the INLA approach (Rue *et al.*, 2009). As mentioned above, the latter is a tool designed to make statistical inference for Latent Gaussian Markov random field models in a computationally efficient way. INLA assumes that a response variable vector $\mathbf{y}$ is observed at a set of n sites and is distributed as a member of the exponential family with possibly vector-valued parameter θ_1

$$E(\mathbf{y}) = d(\mu + \tilde{\mathbf{z}} + \varepsilon), \tag{6}$$
$$\varepsilon \sim N(\mathbf{0}, \tau^2 I),$$

where μ is the intercept, $\tilde{\mathbf{z}}$ is the SPDE representation of a spatial process $\mathbf{z}$ with Matérn covariance function as a GMRF, ε is the vector of n unstructured error terms and τ^2 is the variance of the elements of ε.

Let $\theta_2 = (log(1/\tau^2), log(\kappa^2), log(1/q))'$ be a parameter vector and $\mathbf{x} = (\eta', \mu, \tilde{\mathbf{z}}')' \in R^M$ be a GMRF with assumed zero mean and precision $Q(\theta_2)$ depending on the hyperparameter vector θ_2. The elements of the observed response vector $\mathbf{y}$ are assumed to be conditionally independently distributed given $\mathbf{x}$ and θ_1. The definition of the prior distribution for the hyperparameter vector $\theta = (\theta_1', \theta_2')' \in R^m$ completes the INLA model. In Rue *et al.* (2009) an approximation of the relevant posterior distribution

$$\pi(x_i|\mathbf{y}) = \int \pi(x_i|\theta, \mathbf{y})\pi(\theta|\mathbf{y})d\theta \tag{7}$$

is obtained combining the Laplace approximation of $\pi(x_i|\theta, \mathbf{y})$, $i = 1, \dots, M$ and $\pi(\theta|\mathbf{y})$. The latter is discretized by a grid search and approximate $\pi(\theta_j|\mathbf{y})$, $j = 1, \dots, m$ are used in the numerical approximation of integral (7).

4 Tapering

Tapering speeds up the computational time by introducing zeros in the covariance matrix, thus enabling the use of sparse matrix algorithms when working with a zero mean (without loss of generality) Gaussian process $\mathbf{y} = \mathbf{z} + \varepsilon$ defined on an irregular lattice, where $\mathbf{z}$ is a GF with zero mean and covariance matrix Σ given by an isotropic stationary covariance function $C(h;\theta)$ and ε has i.i.d. elements with variance τ^2. In some applications there is reason to believe that distant sites are independent and then many elements of the spatial covariance matrix Σ can be set to zero. The idea behind tapering is to reduce the covariance to zero for distances greater or equal than a fixed value ρ. This is done by taking an isotropic correlation function on a compact support $C_{tap}(h;\rho)$ that is identically zero if $h \geqslant \rho$.

Let $C(h;\theta)$ be the covariance function of the original process, depending on a vector of parameters θ. The tapered covariance function is:

$$C_t(h;\theta,\rho) = C(h;\theta)C_{tap}(h;\rho), \tag{8}$$

and the associated tapered covariance matrix is:

$$\Sigma^{TAP}(\theta_t) = \Sigma \circ \mathbf{F}(\rho), \tag{9}$$

where $\mathbf{F}(\rho)_{ij} = C_{tap}(h_{ij};\rho)$ and $\circ$ is the Schur product (Hadamard product) that ensures that (9) is a valid covariance matrix (Kaufman *et al.*, 2008). The tapered covariance matrix (9) has thus a large proportion of zero elements and is sparse. The size of ρ defines the strength of the tapering, $\rho \to \infty$ implies no elements of the covariance matrix are forced to zero, while $\rho \to 0$ returns independent observations.

To approximate the Gaussian log-likelihood we can simply replace the original covariance matrix with the tapered covariance:

$$l_{1tap}(\theta_t) \propto -\frac{1}{2}\log\left|(\Sigma+\tau^2 I)\circ \mathbf{F}(\rho)\right| - \frac{1}{2}\mathbf{y}'\left[(\Sigma+\tau^2 I)\circ \mathbf{F}(\rho)\right]^{-1}\mathbf{y} \tag{10}$$

The previous expression is called the *one-taper likelihood approximation*. If the taper range is too small with respect to the correlation range of the process, there is no guarantee that maximizing (10) leads to asymptotically unbiased estimates. A more computationally expensive asymptotically unbiased *two-taper approximation* was also proposed by Kaufman *et al.* (2008). Heuristic work suggests that l_{1tap} is more suitable for the Kriging procedure (Kaufman *et al.*, 2008) than the two-taper.

In Zhang (2004), Zhang and Du (2008) and Kaufman *et al.* (2008), several asymptotic results are given for processes with Matérn covariance function (1) in the fixed domain. These results ensure the identifiability of parameters σ^2 and κ under given conditions and the equivalence of the original process with the process obtained by the tapering. As suggested by Furrer *et al.* (2006), a family of taper function that satisfies the aforementioned conditions is the family of compactly supported functions proposed by Wendland (1995). When the original covariance function is Matérn, two such functions are commonly used:

$$W_{\rho,1}(h) = \max\left[\left(1-\left(\frac{h}{\rho}\right)\right)^4, 0\right]\left(1+4\frac{h}{\rho}\right), \tag{11}$$

$$W_{\rho,2}(h) = \max\left[\left(1-\left(\frac{h}{\rho}\right)\right)^6, 0\right]\left(1+6\frac{h}{\rho}+35\frac{h^2}{2\rho^2}\right). \tag{12}$$

While $W_{\rho,1}$is a valid taper function (i.e the Schur product gives a valid covariance function) for $\nu \leq 1.5$, $W_{\rho,2}$ is valid for $\nu \leq 2.5$ (Bolin and Lindgren, 2011). Furrer *et al.* (2006) report that for $\nu \leq 1.5$ the ratio of the mean squared errors of the $W_{\rho,1}(\cdot)$ tapered process and the original process is closer to 1 than the corresponding ratio obtained with the $W_{\rho,2}(\cdot)$.

5 A simulation study

In this section we are going to discuss the "*big n problem*" in the context of spatial interpolation, comparing SPDE/INLA and tapering approximations when used in a Kriging predictions framework. For this purpose we simulate data from the following model:

$$\begin{aligned} y(\mathbf{s}) &= z(\mathbf{s}) + \varepsilon(\mathbf{s}), \\ z(\mathbf{s}) &\sim N(\mathbf{0}, \Sigma), \\ \varepsilon(\mathbf{s}) &\sim N\left(\mathbf{0}, \tau^2\mathbf{I}\right), \end{aligned} \tag{13}$$

where $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$ is a set of sites, Σ is built according to a spatially structured Matérn covariance function (1) and $\varepsilon(\mathbf{s})$ is a white noise process with variance τ^2 (nugget), then $y(\mathbf{s})$ is an isotropic and stationary process. Even with a large process dimension, simulation from (13) is straightforward and computationally feasible as only one matrix inversion is required and, having chosen a model with measurement error ($\tau^2 \neq 0$), the resulting covariance matrix is numerically stable.

Within this setting we run two types of simulation experiments using the grf function in the geoR R library: a general experiment where 18 datasets are obtained by as many combinations of values of the practical range $\phi = 10, 50, 150$, variance $\sigma^2 = 10, 50$ and nugget $\tau^2 = 2, 5, 10, 25, 50$, keeping $\nu = 1$. The second simulation experiment is focused on a large value of the process variance ($\sigma^2 = 100$), keeping $\nu = 1$. We generate five datasets differing for the values of the practical range and the nugget: one dataset with $\phi = 10$ and $\tau^2 = 20$, one with $\phi = 150$ and $\tau^2 = 20$ and three with $\phi = 50$ and τ^2 equal to 20, 50 and 100. In both experiments, for every combination of covariance structure parameters, we simulate 50000 data points from model (13) with coordinates x and y uniformly generated over the bidimensional interval $[0, 1000] \times [0, 1000]$. Every dataset is divided into a *training* set, denoted by $\mathbf{s}_{tr}$, and a *testing* set, denoted by $\mathbf{s}_{ts}$, obtained by randomly sampling 40000 (training) and 10000 (testing) observations. In the second experiment each simulated dataset is randomly sampled 100 times, in order to build an equivalent number of replications of training and testing sets for every combination of covariance structure parameters. As the focus is on the evaluation of the performance of the approximation returned by the two approaches, we perform simulated experiments in a very simplified setting, keeping the values of covariance structure parameters fixed at the corresponding “true” values, i.e. those used to simulate the data. Also the mean is assumed identically null with the tapering, while an informative prior centered at zero is used with SPDE/INLA. Conditioning on values at the training set, predictions are obtained at the testing set and at $\mathbf{s}_{tr}$ as well.

Root mean squared (prediction) errors (RMSE’s) are computed comparing simulated values at $\mathbf{s}_{tr}$ and $\mathbf{s}_{ts}$ with their corresponding predictions, in order to respectively verify the fitting and predictive ability of the SPDE/INLA and tapering approaches.

5.1 Prediction details: SPDE/INLA

Let us consider a process with $\nu = 1$ in a two-dimensional space. In this setting $\alpha = 2$ and the Gaussian weights in (5) have distribution (Lindgren *et al.*, 2011):

$$\mathbf{w} \sim N(\mathbf{0}, \mathbf{Q}_2^{-1}).$$

As mentioned in section 3, values of the matrices $\mathbf{G}$ and $\tilde{\mathbf{C}}$ required in the calculation of $\mathbf{Q}_2$ are found by the constrained refined Delaunay triangulation (CRDT). First, sites in $\mathbf{s}_{tr}$ and $\mathbf{s}_{ts}$ are used as starting vertices of the triangles. Then extra vertices are added to ensure that the transition between small and large triangles is smooth. Taking a generic triangle T, with vertices $(\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3)$, expressions for matrices $\tilde{\mathbf{C}}(T)$ and $\mathbf{G}(T)$, defined as in section 3 for triangle T, are given by (see Lindgren and Rue, 2007):

$$\tilde{\mathbf{C}}(T) = \frac{|T|}{3} \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}, \tag{14}$$

$$\mathbf{G}(T) = \frac{1}{4|T|} \begin{pmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{r}_3 \end{pmatrix}' \begin{pmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{r}_3 \end{pmatrix}, \tag{15}$$

where $|T|$ is the area of the triangle and

$$\begin{aligned} \mathbf{r}_1 &= \mathbf{t}_3 - \mathbf{t}_2, \\ \mathbf{r}_2 &= \mathbf{t}_1 - \mathbf{t}_3, \\ \mathbf{r}_3 &= \mathbf{t}_2 - \mathbf{t}_1 \end{aligned}$$

are the edge vectors opposite to each corner, e.g $\mathbf{r}_1$ is the edge opposite to $\mathbf{t}_1$.

Once matrices (14) and (15) are obtained for all triangles, $\mathbf{G}$ and $\tilde{\mathbf{C}}$ are easily computed. For $i = 1, \ldots, N$ let T_i be the set of triangles linked to vertex $\mathbf{t}_i$, then the i-th element of the diagonal of matrix $\tilde{\mathbf{C}}$ is:

$$\tilde{\mathbf{C}}_{i,i} = \langle \psi_i, 1 \rangle = \sum_{T \in T_i} \tilde{\mathbf{C}}_{i,i}(T) = \frac{1}{3} \sum_{T \in T_i} |T|, \tag{16}$$

and the ik-th element of $\mathbf{G}$ is

$$\mathbf{G}_{i,k} = \langle \nabla \psi_i, \nabla \psi_k \rangle = \sum_{T \in T_i \cap T_k} \mathbf{G}_{i,k}(T). \tag{17}$$

In this study three SPDE specifications are applied to each dataset, they differ by the minimum angle permitted in the triangulation algorithm: 10, 15 and 20 degrees respectively. The three models are named Spde10, Spde15 and Spde20. Notice that by increasing the minimum angle we indirectly increase the number of extra vertices used, this allows to obtain more precise results at a higher computational cost. As mentioned above, we obtain SPDE/INLA predictions conditioning on the values of ν, κ^2, q and τ fixed at the corresponding values used to simulate the data and we adopt a Gaussian prior with mean zero and variance 0.01 for the mean. These computations are implemented within the INLA package in the R system.

5.2 Prediction details: tapering

When the tapering is used with Kriging it allows to reduce the computational effort needed to invert covariance matrices involved in spatial interpolation. Let $\mathbf{s}_{pr}$ be a set of points where we want to make predictions and Σ_{tr}, Σ_{pr} and $\Sigma_{pr,tr}$ be the covariance matrices of z at $\mathbf{s}_{tr}$, $\mathbf{s}_{pr}$, and between $\mathbf{s}_{pr}$ and $\mathbf{s}_{tr}$ respectively. We obtain:

$$z(\mathbf{s}_{pr})\,|y(\mathbf{s}_{tr}) \sim N\left(\mu_{pr|y}, \Sigma_{pr|y}\right), \tag{18}$$

where

$$\begin{aligned} \mu_{pr|y} &= \Sigma_{pr,tr}\left(\Sigma_{tr} + \tau^2 \mathbf{I}\right)^{-1} y(\mathbf{s}_{tr}), \\ \Sigma_{pr|y} &= \Sigma_{tr} - \Sigma_{pr,tr}\left(\Sigma_{tr} + \tau^2 \mathbf{I}\right)^{-1} \Sigma'_{pr,tr}, \end{aligned} \tag{19}$$

Expressions in (19) involve matrix inversions that can cause troubles when n is large. In the one-taper approach the two matrices $\Sigma_{pr,tr}$ and Σ_{tr} are replaced with their tapered versions $\Sigma_{pr,tr}^{TAP}$ and Σ_{tr}^{TAP} obtained as:

$$\begin{aligned} \Sigma_{pr,tr}^{TAP} &= \Sigma_{pr,tr} \circ \mathbf{F}_{pr,tr}(\rho), \\ \Sigma_{tr}^{TAP} &= \Sigma_{tr} \circ \mathbf{F}_{tr}(\rho), \end{aligned} \tag{20}$$

where $\mathbf{F}_{tr}$ and $\mathbf{F}_{pr,tr}$ are calculated considering distances among locations in the training set for $\mathbf{F}_{tr}$ and between locations in $\mathbf{s}_{pr}$ and $\mathbf{s}_{tr}$ for $\mathbf{F}_{pr,tr}$. The tapered versions of the Kriged mean and variance in (19) become:

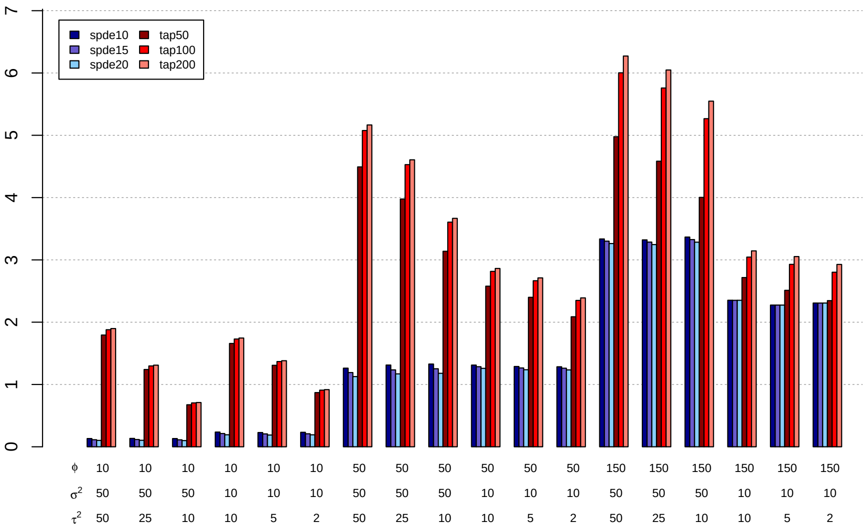

Figure 1: RMSE's calculated for the training sets in the first simulation experiment. Parameter combinations used for data simulation are indicated under the horizontal axis ($\mu = 0$, $\nu = 1$), the legend contains SPDE/INLA and tapering specifications.

$$\begin{aligned}\mu^{TAP}_{pr|y} &= \Sigma^{TAP}_{pr,tr}\left(\Sigma^{TAP}_{tr}+\tau^2\mathbf{I}\right)^{-1} y\left(\mathbf{s}_{tr}\right),\\ \Sigma^{TAP}_{pr|y} &= \Sigma^{TAP}_{tr} - \Sigma^{TAP}_{pr,tr}\left(\Sigma^{TAP}_{tr}+\tau^2\mathbf{I}\right)^{-1}\left(\Sigma^{TAP}_{pr,tr}\right)'.\end{aligned} \tag{21}$$

For each dataset in our experiments we use three tapering specifications with $\rho = 50, 100, 200$, respectively named Tap50, Tap100 and Tap200. Notice that larger values of ρ imply less sparse covariance matrices and larger neighborhoods for carrying on the Kriging interpolation, with corresponding higher computational costs. As with the SPDE/INLA, we consider covariance structure parameters as fixed and known and $\mu = 0$. The tapered Kriging predictions are obtained with the R function mKrig in package fields.

5.3 Results

In the following we compare the approximation performance of SPDE/INLA and tapering, running a general simulation experiment and a second simulation experiment focused on a large value of the process variance. Root mean squared (prediction) errors are computed in order to verify the fitting and predictive ability of the two approaches.

First Experiment: Fig. 1 describes RMSE's calculated for the training sets and shows some interesting features of the fitting ability of the two approaches: the SPDE/INLA approach has a better overall fit

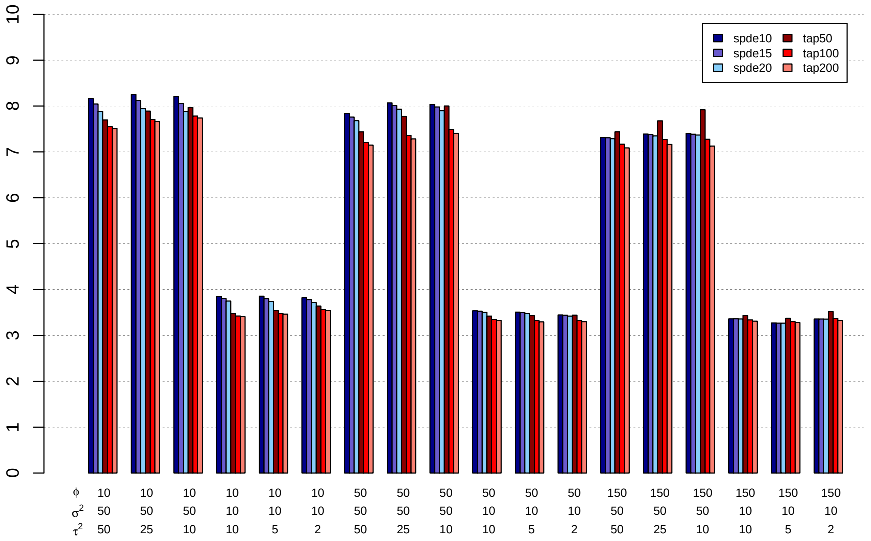

Figure 2: RMSE's calculated for the testing sets in the first simulation experiment. Parameter combinations used for data simulation are indicated under the horizontal axis ($\mu = 0$, $\nu = 1$), the legend contains SPDE/INLA and tapering specifications.

than the tapering and is also less sensitive to the presence of noise on equal terms. While τ^2 is well identified in the SPDE, tapering suffers from variance component lack of identifiability and from matrix ill-conditioning that arises in spatial models with small or null noise variance (Banerjee *et al.*, 2004). Both methods improve when ϕ decreases as a GF with weaker spatial correlation is less affected by the approximation. SPDE/INLA has a better fit when σ^2 decreases and the range is large ($\phi = 150$), the same holds for tapering also with a smaller value of the range ($\phi = 50$). Smaller values of the tapering parameter improve the fit as predictions at training locations are based on a more "local" covariance structure.

The predictive ability of the two approaches is represented in Fig. 2 that shows huge differences for both methods according to the σ^2 specification, with better predictions for smaller variances. SPDE/INLA produces more accurate predictions for larger values of the range when the variance is small. Overall tapering seems to return more accurate predictions than the SPDE/INLA when the process range ϕ is small, even when the tapering parameter is large. If ρ is widely underestimated with respect to the range value, the prediction accuracy of the tapered estimates is strongly reduced. However when the spatial dependence becomes stronger (larger ϕ) the two approximations return an equivalent prediction accuracy when the tapering parameter gets closer to the process range. Finally, for small values of τ^2, ill-conditioned covariance matrices hamper the predictive ability of the tapering method.

Second Experiment: Smaller RMSE's are observed for the training sets (better fit) of all SPDE/INLA specifications with respect to the corresponding tapering ones (Tab. 1). Both testing and training sets, in the first and second experiment show the same behavior with respect to the SPDE specification: the values

Table 1: Average RMSE's computed for 100 training sets in the second simulation experiment. Parameter combinations used for data simulation on rows ($\mu = 0$, $\nu = 1$, $\sigma^2 = 100$), SPDE/INLA and tapering specifications on columns.

Process parameters		Spde10	Spde15	Spde20	Tap50	Tap100	Tap200
ϕ	τ^2						
10	20	0.106	0.089	0.078	0.585	0.609	0.615
50	20	1.201	1.106	1.025	3.428	3.971	4.039
50	50	1.223	1.130	1.046	4.727	5.407	5.497
50	100	1.165	1.079	1.001	5.474	6.244	6.354
150	20	3.644	3.565	3.482	4.825	6.673	7.046

Table 2: Average RMSE's computed for 100 testing sets in the second simulation experiment. Parameter combinations used for data simulation on rows ($\mu = 0$, $\nu = 1$, $\sigma^2 = 100$), SPDE/INLA and tapering specifications on columns.

Process parameters		Spde10	Spde15	Spde20	Tap50	Tap100	Tap200
ϕ	τ^2						
10	20	11.642	11.407	11.184	11.411	11.127	11.062
50	20	11.419	11.309	11.152	11.479	10.663	10.534
50	50	11.493	11.370	11.210	11.111	10.408	10.289
50	100	11.239	11.133	10.990	10.662	10.112	10.010
150	20	10.420	10.383	10.334	11.242	10.173	9.938

of the RMSE decrease with a more complex triangulation, thus both the fit (RMSE's in Fig. 1 and Tab. 1) and the predictive ability (RMSE's in Fig. 2 and Tab. 2) of the method improve with a growing number of vertices. A different behavior is observed for the tapering in the two sets: in the training sets the RMSE is inversely proportional to ρ (a small tapering parameter produces smaller RMSE's) while the opposite holds in the testing sets (the RMSE values decrease when ρ increases). A small value of ρ produces predictions at training points being based on few neighboring points (overfitting), leading to larger discrepancies with the testing set and larger values of the corresponding RMSE's. While for the testing sets (Tab. 2) the tapering specifications $\rho = 100, 200$ have always smaller RMSE's (better predictive ability) than the three SPDE/INLA specifications, when the nugget is small ($\tau^2 = 20$) Tap50 only outperform Spde10 if the practical range equals 10, i.e. when the spatial correlation is very weak. As already reported, small values of the nugget result in less accurate tapering predictions. For larger values of τ^2, differences in favor of the tapering predictive performance become more evident.

6 Concluding remarks

In this work we have offered a brief overview of recent developments in spatial modeling for large datasets. Challenges are there to solve the "*big n problem*". Extensions to complex setting such as multivariate spatial data or space-time data are still in progress, as the adaptation of SPDE approximation of GF in multivariate settings, while low rank approximations already offer promising solutions in this regard (Banerjee and Fuentes, 2012).

Here we focused mostly on two approaches, SPDE/INLA and tapering that are based on different philosophies while suitable for the same type of data. In the SPDE/INLA a rigorous solution to the approximation of a GF with GMRF is found and then the GMRF estimation and prediction is carried on in a Bayesian perspective using a non-McMC method. The latter adds a further approximation level (introduced by the INLA) to the involved distributions. Tapering introduces only one level of approximation directly on the second moment of the process (the covariance matrix). In this case estimation is carried on in a likelihood framework, while performing a full Bayesian estimation with the tapering approach does not seem straightforward as pointed out in Sun *et al.* (2012); a quasi-Bayesian procedure has been recently proposed by Shaby and Ruppert (2011) were the full likelihood is substituted by a tapered likelihood. In our opinion SPDE/INLA is a better choice when a Bayesian implementation is preferred, while tapering seems to be best used when maximum likelihood is the choice.

In our simulation experiments both the fitting and predictive ability show the same behavior with respect to the SPDE/INLA specification: the values of the RMSE decrease with a more complex triangulation. A different behavior is observed for the tapering: the fit is inversely proportional to ρ (a small tapering parameter produces smaller RMSE's) while the opposite holds for the predictive ability (RMSE values decrease when ρ increases). Notice that small values of the nugget result in less accurate tapering predictions.

As we mentioned in the introduction, no one of the methods proposed so far to approach "*the big n problem*" is "good" for all estimation situations. The SPDE/INLA and the tapering, however, are suitable whenever isotropic stationary Kriging has to be performed. In a more complex setting such as space-time modeling, SPDE/INLA is easily adapted, while tapering needs the assumption of separability (Sun *et al.*, 2012). For non stationary fields, as mentioned in Section 3, Lindgren *et al.* (2011) propose a solution to the SPDE with q and κ^2 varying in space, while the tapering approach can be used but its behavior and accuracy are not fully understood yet (Sun *et al.*, 2012).

References

Banerjee, S. and Fuentes, M. Bayesian modeling for large spatial datasets. *WIREs Computational Statistics*, **4**, 59–66 (2012).

Banerjee, S., Carlin, B. P., and Gelfand, A. E. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall (2004).

Banerjee, S., Gelfand, A., Finley, A., and Sang, H. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 825–848 (2008).

Bolin, D. and Lindgren, F. Spatial wavelet Markov models are more efficient than covariance tapering and process convolutions. *arXiv:1106.1980v1* (2011).

Brenner, S. C. and Scott, R. *The Mathematical Theory of Finite Element Methods*. Springer (2007).

Cressie, N. and Johannesson, G. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 209–226 (2008).

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

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

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

Haas, T. Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *Journal of the American Statistical Association*, **90**, 1189–1199 (1995).

Higdon, D., Swall, J., and Kern, J. Non-stationary spatial modeling. *Bayesian Statistics*, **6**, 761–768 (1998).

Ji, W. Y., Simon, W., Koray, K., and Ercan, E. K. Variant functional approximations for latent Gaussian models. Technical Report of Statistics Department, Trinity College Dublin (2011).

Kaufman, C., Schervish, M., and Nychka, D. Covariance tapering for likelihood based estimation in large spatial data set. *Journal of the American Statistical Association*, **103**, 1545–1555. (2008).

Lindgren, F. and Rue, H. Explicit construction of GMRF approximations to generalised Matérn fields on irregular grids. *Scandinavian Journal of Statistics*, **35**, 691–700 (2007).

Lindgren, F., Rue, H., and Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: the Stochastic Partial Differential Equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498 (2011).

Mardia, K., Goodall, C., Refern, E., and Alonso, F. The kriged Kalman filter. *Test*, **7**(217–285) (1998).

Matsuda, Y. and Yajima, Y. Fourier analysis of irregularly spaced data on R^d. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 191–217 (2009).

Rubinstein, B. Y. *Simulation and the Monte Carlo Method*. Wiley and Sons, New York (1981).

Rue, H. and Held, L. *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall, 1st edition (2005).

Rue, H. and Tjelmeland, H. Fitting Gaussian Markov Random Fields to Gaussian Fields. *Scandinavian Journal of Statistics*, pages 31–49 (2002).

Rue, H., Martino, S., and Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 319–392 (2009).

Sampson, P. and Guttorp, P. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, **87**, 108–119 (1992).

Shaby, B. and Ruppert, D. Taper covariance: Bayesian estimation, asymptotics, and applications. *Journal of Computational and Graphical Statistics*, **to appear** (2011).

Stein, M. L., Chi, Z., and Welty, L. J. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**, 275–296 (2004).

Sun, Y., Li, B., and Genton, M. G. (2012). Geostatistics for large datasets,. In J. Montero, E. Porcu, and M. Schlather, editors, *Advances And Challenges In Space-time Modelling of Natural Events*, volume 207 of *Lecture notes in Statistics*, chapter 3, pages 55–77. Springer.

Wendland, H. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, **4**, 389–396 (1995).

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

Whittle, P. Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute*, **40**, 974–994 (1963).

Zhang, H. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**, 250–261 (2004).

Zhang, H. and Du, J. Covariance tapering in spatial statistics. In J. Mateu and E. Porcu, editors, *Positive definite functions: From Schoenberg to space-time challenges*. Gráficas Castañ, s.l. (2008).