Non-stationary spatial distribution models, and its application in marine megafauna.

Martina Le-Bert Heyl

2024-01-28

Contents

Introduction

The availability of spatial methods in statistics has increased substantially in the past decades due to advances in sampling tools as-well-as computational efforts to analyze it. Some examples include data collected from drones, and satellite images (**??**), and software development like arcgis (**?**) or specialized R packages (**?**). Spatial analysis has become essential for many intriguing questions in fields ranging from epidemiology to ecology (**???**). However, non of it comes without its challenges, learning from the data requires statistical models that can resemble reality. This is of particular importance when studying phenomena or patterns at a local scale.

Models used then should allow researchers to take into account all available information including spatial correlation. Many ecological datasets exhibit spatial correlation in the observed variables due to biotic or abiotic processes, such as dispersal limitation, social aggregation, and spatial structure in unobserved explanatory variables. Whether the observations are points in space (e.g., the location of an endangered species), counts (e.g., the number of individuals of the endangered species in the sample space), or values of some continuous variable (e.g., water nutrient levels at sampled points), spatial correlation means that each observation depends on any other observation within some unknown range of correlation. Unlike models assuming independence between observations, Bayesian inference models allow us to consider spatial correlation.

For this reason, the implementation of spatially explicit Bayesian models has received increasing attention. In other words, it is becoming more and more necessary for researchers to use Bayesian inference as a tool to deal with georeferenced data.

In this introductory section a brief review on the following is given, i) Bayesian theory and computation; ii) Integrated Nested Laplace Approximation (INLA); iii) Stochastic Partial Differential Equations (SPDE), and iv) point processes. Although the entirety of the project doesn't intend to profoundly do research on these topics a preliminary discussion on it is necessary to understand the proposed model and applications later on.

Bayesian Theory and computation

A starting point in Bayesian analysis is the consideration of parameters θ as random variables in contrast to the frequentist perspective, which treats parameters as fixed but unknown values. Then, knowledge about θ comes from the distribution of θ given the data y. Getting this distribution is done using Bayes theorem.

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)}$$

In the equation, $\pi(\theta|y)$ is the *posterior* distribution of θ given the data y; $\pi(y|\theta)$ is the likelihood of the data y given θ; $\pi(\theta)$ is *prior* density function for θ; and $\pi(y)$ is the marginal likelihood of y that acts as a normalizing constant. The latter is often hard to obtain so Bayes theorem is simplified to,

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$$

In the previous equations $\pi(\theta|y)$ is actually the *joint posterior* distribution. To obtain individual an parameter the *marginal posterior* for that parameter has to be calculated. This is done by integrating θ_{-i} parameters out, θ_i being the parameter of interest.

$$\pi(\theta_i|y) = \int \pi(\theta|y) d\theta_{-i}$$

with, $i = 1, ..., dim(\theta)$.

Bayesian theory is clear, but the calculation of posterior densities can be computationally intensive. The following describes random sampling and approximation methods for computation in Bayesian statistics.

Sampling: The goal of methods based on random sampling is to generate and sample posterior distributions arbitrarily. The main sampling methods are based on generating a Markov chain whose stationary distribution is the posterior distribution. Collectively, these methods are known as Markov Chain Monte Carlo (MCMC) (**??**). Considering the differences among algorithms implementing the MCMC method, including the Metropolis-Hastings, Gibbs, and Hamiltonian Monte Carlo algorithms (**???**), in general:

1. Specify the model and chosen prior distribution.
2. Specify the initial values of the chain.
3. Let the chain run until it reaches a stationary state. All observations before this point are discarded as warm-up values.
4. Obtain the posterior distribution from the sampling.

For each step, there are certain constraints that can make this analysis quite complex and extensive. For example, observations from the chain are positively correlated, so the sampling from the posterior distribution generated by the chain must be large enough to discard dependence. The fact that the posterior distribution comes from sampling means that a sufficient number of samples must be generated for the estimation to be accurate, which can be time-consuming.

Approximation: These methods originated with the article by Tierney and Kadane (**?**). Among them, but more recent, is the method of Integrated Nested Laplace Approximation (INLA) (**?**). The main idea in this case is to obtain the marginal posterior of the parameters $\pi(\theta_i|y)$ instead the joint posterior.

INLA

Three key components are required by INLA: the latent Gaussian models (LGM) framework, a Gaussian Markov random field (GMRF) (**?**), and the Laplace approximation. A brief explanation of these followed by the INLA-SPDE method is given here (**??**).

LGM

A three-stage hierarchical model formulation is obtained as follows:

$$\mathrm{y} \mid \mathrm{x}, \theta_1 \sim \prod_{i \in \mathcal{J}} \pi\left(y_i \mid x_i, \theta_1\right)$$

In this first equation x is a Gaussian random field, θ_1 hyperparameters, and y independent of observations given x, and θ_1.

$$\mathrm{x} \mid \theta_2 \sim \mathcal{N}\left(\mu\left(\theta_2\right), \mathrm{Q}^{-1}\left(\theta_2\right)\right)$$

In the second equation θ_2 are hyperparameters for the latent GMRF structure of the model,

$$\pi(\mathrm{x}, \theta \mid \mathrm{y}) \propto \pi(\theta) \pi(\mathrm{x} \mid \theta) \prod_{i \in \mathcal{J}} \pi\left(y_i \mid x_i, \theta\right)$$

Finally, in the third equation $\theta = (\theta_1, \theta_2)$. This equation comes from $\pi(\mathbf{x}, \theta \mid \mathbf{y}) \propto \pi(\mathbf{x}, \theta) \pi(\mathbf{y} \mid \mathbf{x}, \theta)$, where the joint distribution $\pi(\mathbf{x}, \theta)$ is rearrange as $\pi(\theta)\pi(\mathrm{x} \mid \theta)$, and $\pi(\mathbf{y}|\mathbf{x}, \theta)$ as $= \prod_{i \in \mathcal{J}} \pi\left(y_i \mid x_i, \theta\right)$.

There is a few assumptions needed to be considered for the vector of hyperparameters . 1) the number of hyperparameters is small; 2) as mentioned, the distribution of the latent field $\mathbf{x}|$ is a GMRF; 3) the observations $\mathbf{y}$ are mutually

conditional independent of $\mathbf{x}$ and . These are assumptions for LGMs also hold for the generalized linear models setup described next.

LGMs can be placed in a generalized linear models setup. The likelihood $\pi(y_i \mid x_i, \theta)$ is interpreted so that y_i only depends on its linear predictor x_i, then $\{x_i, i \in \mathcal{I}\}$ is interpreted as the linear predictor η_i,

$$\eta_i = \mu + \sum_j \beta_j z_{ij} + \sum_k f_{k,j_k(i)}$$

where μ is the overall intercept, $\mathbf{z}$ fixed covariates with linear effects β_j, and f_k model components in which j contributes to the ith linear predictor. *()* give examples of model components f_k and its assumptions.

When Gaussian priors for the intercept and the parameters of the fixed effects are assumed, the joint distribution of x is Gaussian,

$$\mathrm{x} = (\eta, \mu, \beta, \mathrm{f}_1, \mathrm{f}_2, ...)$$

x is the latent field in the LGM formulation meaning it has a GMRF structure, and observations are considered independent under $\mathbf{x}$. In this setup the vector of hyperparameters contains the parameters from the likelihood and model components.

GMRF

A GMRF is a multivariate Gaussian random variable with additional properties of conditional independence (also called the "Markov property"): x_i and x_j are conditionally independent given the remaining elements if and only if the entry (i, j) of the precision matrix is 0. The sparsity of the precision matrix is a key property for computational efficiency so the latent field should not only be Gaussian but also a GMRF for the INLA method to be efficient.

As LGMs, GMRF can also be constructed additively. As mentioned before $\mathbf{x}$ is GMRF with precision matrix as result of the sum of the precision matrices of the fixed effects and model components.

Besides counting with sparse matrices, treating $\mathbf{x}$ as a GMRF has also advantages in the approximation of this joint posterior distribution.

Laplace Approximation

The Laplace approximation is used to approximate the integral $I_n = \int_x \exp(nf(x))dx$ as $n \to \infty$ by doing a Taylor expansion on $f(x)$ at its mode x_0.

$$
\begin{aligned}
I_n &\approx \int_x \exp\left(n\left(f(x_0) + (x - x_0) f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(x_0)\right)\right) dx \\
&= \exp(nf(x_0)) \int \exp\left(\frac{n}{2}(x - x_0)^2 f''(x_0)\right) dx \\
&= \exp(nf(x_0)) \sqrt{\frac{2\pi}{-nf''(x_0)}} = I_n
\end{aligned}
$$

with $f'(x_0)$ and $f''(x_0)$ the first and second order derivatives at x_0. Then, I_n is the Gaussian integral and $f'(x_0) = 0$. If $nf(x)$ is taken to be the sum of log-likelihoods, x the unknown parameter, and the central limit theorem holds, the approximation will be exact as $n \to \infty$.

In INLA this approximation is used to some of the posterior distributions, for example joint posterior distribution of the hyperparameters $\pi(\theta \mid \mathbf{y})$.

SPDE

Gaussian random fields (GRFs) are a widely used tool for capturing dependency in spatial modeling. GRFs have the advantage of being fully characterized by its mean and covariance, the latter has a parametrized structure often from the Matérn covariance family, with the parameters termed hyperparameters. However, some computational challenges have to be overcome to estimate the covariance matrix . With INLA, computational speed and accuracy comes from the sparse precision matrix representation $\mathbf{Q}$. As mentioned before the precision matrix is the inverse of the covariance matrix, then $\mathbf{Q} = {}^{-1}$. In the spatial statistics context sparsity is attained using the SPDE approach.

The SPDE approach constructs a continuously indexed approximation of the GRF by using an SPDE model instead of a discrete model on a set on a grid. The following is a brief summary on the intuition behind the SPDE approach.

Matérn covariance function and SPDE

The use of a Matérn covariance function is convenient as its parameters have a clear physical interpretation. However, covariance matrices calculations are not computationally efficient. The Matérn covariance function is given by,

$$
Cov\left(x(s_i), x(s_j)\right) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{8\nu}\left\|s_i - s_j\right\| / \rho\right) K_\nu \left(\sqrt{8\nu}\left\|s_i - s_j\right\| / \rho\right)
$$

where $||s_i - s_j||$ is the Euclidean distance between points s_i and s_j, σ^2 is the marginal standard deviation, ν is the smoothness parameter, K_ν is the modified

Bessel function of the second kind and order $\nu > 0$, and ρ is the spatial distance at which correlation is approximately 0.1.

In parallel, results from Whittle (**?**) show that the stationary solution of an SPDE of the following form

$$(\kappa^2 - \Delta)^{\alpha/2} (\tau x(s)) = \mathcal{W}(s), \quad s \in \mathbb{R}^d$$

has a Matérn covariance function representation. In the equation, $x(s)$ is a GRF, $\Delta = \sum_i \partial^2 / \partial s_i^2$ is the Laplacian operator and $\mathcal{W}$ is some Gaussian white noise. Parameters $\kappa > 0$, and $a > d/2$ are scale, and smoothness parameters, and $\tau > 0$ controls the variance.

Parameters used in the two previous equations are not the same, however there is a one-to-one correspondence between them. In other words, provided some restrictions on the SPDE, there is an approximation to the solution that is continuously indexed and has the Matérn covariance structure approximately.

FEM

Without getting into too much detail the Finite Element Method (FEM) is employed to solve the Stochastic Partial Differential Equation (SPDE), which allows solutions to be approximated in an irregular grid. In this case the spatial domain is divided into a set of non-intersecting triangles creating a triangulated mesh with n nodes or triangle vertices. For each mesh node there is a basis function Ψ_k which is in turn a piece wise linear function. The approximation is as follows,

$$x(\mathbf{s}) = \sum_{k=1}^{n} \psi_k(\mathbf{s}) w_k$$

Here, the continuously indexed GRF $x(s)$ is represented as a (discretely indexed) GMRF by a sum of ψ_k basis functions and w_k Gaussian distributed weights, with $k = 1, ..., n$. The solution preserves the sparsity of the precision matrix for the random field at mesh nodes by carefully choosing the basis functions.

Projection matrix

Now, from the implementation point of view a projection matrix $\mathbf{A}$ is used in INLA to project observations to the triangles nodes in the mesh. Each row in $\mathbf{A}$ corresponds to an observation i at location s_i, and each column to a mesh node. Values in $\mathbf{A}$ are given by the location s_i of the observation in the mesh. If the location s_i lies inside a triangle the ith row of $\mathbf{A}$ will have three non-zero entries in the columns of $\mathbf{A}$ corresponding to the three nodes that construct the

triangle where s_i lies in. If s_i lies along the edge of a triangle there will be two non-zero entries in the ith row of $\mathbf{A}$; and if s_i is on top of a triangle node there will be just one non-zero entry. Finally each row in $\mathbf{A}$ sums up to 1.

Finally, since $x(s)$ is a GRF, a vector of any set of locations will be a multivariate Gaussian vector. The vector's covariance matrix according to the SPDE approximation is,

$$= \mathbf{A}\mathbf{Q}^{-\mathbf{1}}\mathbf{A}^{\mathbf{T}}$$

As mentioned in the beginning, this section only intends to give some broad intuition behind Bayesian computation and INLA methods. A detailed explanation can be found in **?**; **?**; **?**; **?**; **?**; **?**.

Point processes

A point pattern refers to the arrangement or distribution of individual points or locations within a study area. These points represent the sampled elements or observations in the context of spatial data. Point patterns are commonly encountered in various fields, such as ecology, epidemiology, forestry, and geography, where the goal is to understand the spatial distribution of events or phenomena.

In point processes modelling it is assumed the occurrence of observed events is influenced by an inherent spatial process. The underlying process is modeled using an intensity function $\lambda(s)$, which it is taken to be the intensity of a Poisson process.

Particularly, a Log-Gaussian Cox Process (LGCP) is a statistical model used for analyzing the distribution of events or occurrences in continuous space.

More in detail, the Cox process is a type of point process that describes the occurrence of events over a continuous spatial domain. It is characterized by an intensity function, which represents the expected number of events per unit area.

The following equation illustrates the number of events in a region A follow a Poisson process with intensity λ_A.

$$\lambda_A = \int_A \lambda(s)ds$$

In the LGCP, the intensity function of the Cox process is log-transformed. Conditional on a set of hyper-parameters, the logarithm of the intensity function is then assumed to follow a Gaussian distribution.

The Gaussian process is used to model the spatial variation in the log-intensity function. It is a stochastic process where the values at any finite set of locations are jointly Gaussian distributed. Then, $log\lambda(s) = Z(s)$, and $Z(s)$ a Gaussian process. Moreover, an LGCP is an inhomogeneous Poisson process, conditional on a realization of $Z(s)$

Additionally, **?** developed a method for fitting LGCP models considering an SPDE approach. Because of it it is possible to fit an LGCP model over a mesh instead of a regular grid.

Chapter 1

Transparent Barrier model

1.1 Motivation

Traditional spatial statistics rely on stationary models like the Matérn field, but they can smooth over features such as boundaries, holes, or physical barriers. For such cases **?** proposed the Barrier model as a solution which differs from established methods in the sense that it doesn't rely on the shortest distance between two points for the construction of the covariance matrix function. The Barrier model is based on interpreting the Matérn correlation as a collection of paths through a Simultaneous Autoregressive (SAR) model, manipulating local dependencies to cut off paths crossing physical barriers and formulated as a stochastic partial differential equation (SPDE) for well-behaved discretization.

Until now only *permanent physically impermeable* barriers have been considered in the model. However this turns out to be quite restrictive since some barriers tend to have some degree of *transparency* instead of being permanently impermeable. So a Transparent Barrier model is proposed when barriers with different transparency levels need to be included in the same model. To illustrate this, one can think of some fish that leaves in shallow sea waters. An impermeable barrier would be a set of islands where there is no scenario in which fishes go over it - *we will refer to this as 0 transparency later on* -. However, there might be sand patches with varying water coverage depending on the tide. These sand patches cannot be considered *permanently impermeable* barriers as fishes will be present, but will do so less often than in the *normal* non barrier area. Then, there is the need of counting with a model that can include both barriers and their specific transparency level as some parameter in the model. Moreover, the Barrier model is the case where all the barrier areas have the same transparency level equal to 0 in the Transparent Barrier model framework.

The main advantage of the Transparent Barrier model is it can handle spatial structures with complex barriers of different nature while keeping the computa-

tional efficiency of the stationary model. Next chapters also show Transparent Barrier models are pertinent in real life applications since complex physical barriers are encountered often in spatial modeling.

1.2 Barrier model background

Spatial Gaussian fields (SGFs) are typically assumed to be stationary and isotropic (from here on both qualities are referred to as stationarity), implying that the model component remains unchanged when the underlying map is moved or rotated.

However, the application of stationarity becomes unrealistic in the presence of physical barriers or holes in the study area, challenging the assumption that changing the map's position or orientation shouldn't affect the model, especially when considering the impact of these features on spatial dependencies. The Barrier model proposed in **?** offers a solution in case of non-stationarity SGFs within the INLA-SPDE framework described in the previous section.

To understand the intuition behind the Barrier model only the re-parametrization of the Matérn covariance function and the SPDE approximation are shown next. To get all the details that follow, see **?**.

To start with, the re-parametrization of the Matérn covariance function introduces an r parameterm, which we will need in the Transparent Barrier model framework too.

Let the following be Matérn covariance function,

$$Cov\left(x(s_i), x(s_j)\right) = \sigma_x^2 \left(d\sqrt{8\nu}/\rho\right) K_\nu \left(d\sqrt{8\nu}/\rho\right)$$

where $d = ||s_i - s_j||$ is the Euclidean distance between points s_i and s_j, σ_x^2 is constant, ν is the smoothness parameter, K_ν is the modified Bessel function of the second kind and order $\nu > 0$, and ρ is the spatial distance at which correlation is approximately 0.1.

Now, the re-parametrized Matérn covariance function is,

$$Cov\left(x(s_i), x(s_j)\right) = \sigma_x^2 \left(\frac{d\sqrt{8}}{r}\right) K_\nu \left(\frac{d\sqrt{8}}{r}\right)$$

where $r = \rho/\sqrt{8}$ is the range parameter. r is constant and interpretable because the correlation between two points that are r units apart is near 0.1.

The SPDE is also re-parametrized to get the following,

$$x(s) - \nabla \cdot \frac{r^2}{8} \nabla x(s) = r\sqrt{\frac{\pi}{2}} \sigma_x \mathcal{W}(s)$$

where $x(s)$ the Gaussian field for $s \in \Omega \subseteq \mathbb{R}^2$, $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$, and $\mathcal{W}(s)$ white noise.

What follows is to introduce a different Matérn field for *normal* area and *barrier* area. Both have the same σ but different r. r for the barrier area is close to 0 to remove the correlation in this area. The following are SPDE representations for both Matérn field,

$$x(s) - \nabla \cdot \frac{r_n^2}{8} \nabla x(s) = r_n \sqrt{\frac{\pi}{2}} \sigma_x \mathcal{W}(s), \text{ for } s \in \Omega_n$$
$$x(s) - \nabla \cdot \frac{r_b^2}{8} \nabla x(s) = r_b \sqrt{\frac{\pi}{2}} \sigma_x \mathcal{W}(s), \text{ for } s \in \Omega_b,$$

where r_n is the range parameter for the normal area, r_b is the range parameter for the barrier area and its a fixed fraction of r so that $r_b = r_n/h$, where h is a constant (e.g. 10). Ω_n the normal area, and Ω_b the barrier area. The disjoint union of both Ω_n and Ω_b gives the whole study area Ω.

For details on this re-parametrization, and conditions that need to be satisfied see **?**; **?**; **?**.

1.3 Transparent Barrier model

The Transparent Barrier model approach comes directly from the barrier model. The difference is in thinking of the study area Ω not only as Ω_n and Ω_b, but Ω_n and Ω_{b_i} with $i = 1, ..., l$ and l the number of different transparency levels considered. Then the SPDE representations for the Matérn field for the normal area and the Matérn field for each barrier area with different transparency level,

$$x(s) - \nabla \cdot \frac{r_n^2}{8} \nabla x(s) = r_n \sqrt{\frac{\pi}{2}} \sigma_x \mathcal{W}(s), \text{ for } s \in \Omega_n$$
$$x(s) - \nabla \cdot \frac{r_{b_1}^2}{8} \nabla x(s) = r_{b_1} \sqrt{\frac{\pi}{2}} \sigma_x \mathcal{W}(s), \text{ for } s \in \Omega_{b_1}$$
$$x(s) - \nabla \cdot \frac{r_{b_2}^2}{8} \nabla x(s) = r_{b_2} \sqrt{\frac{\pi}{2}} \sigma_x \mathcal{W}(s), \text{ for } s \in \Omega_{b_2}$$
$$\vdots$$
$$x(s) - \nabla \cdot \frac{r_{b_l}^2}{8} \nabla x(s) = r_{b_l} \sqrt{\frac{\pi}{2}} \sigma_x \mathcal{W}(s), \text{ for } s \in \Omega_{b_l}$$

where $x(s)$ the Gaussian field, $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$, and $\mathcal{W}(s)$ white noise. r_n is the range parameter for the normal area, and r_{b_1} to r_{b_l} the range parameter for the barriers with different transparency level. Also $r_{b_i} = r_n/h_i$, where $h_i \geq 1$ is