Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 67-75

Challenges in Flexible and Scalable Modeling of Point-referenced Spatial Data

Suman Guha
Department of Statistics, Presidency University, Kolkata

Received: 15 June 2025; Revised: 25 June 2025; Accepted: 27 June 2025

Abstract

Statistical analysis of point-referenced spatial/geostatistical data generally considers a multivariate Gaussian distribution as the underlying probability model. That way, the related statistical inference boils down to estimating the mean vector and the covariance matrix of some multivariate normal distribution. While a fully general specification of the covariance matrix yields a flexible model for the data, it introduces too many parameters for consideration, thereby rendering statistical inference impossible. Alternatively, one can use a parametric covariance function that aligns with the underlying data. This covariance function is then used to form the elements of the covariance matrix under consideration. Parametric covariance functions often rely on the assumption of isotropy, or if not so, at least assume stationarity. However, stationary covariance functions are inadequate for explaining the complex dependence structure of spatial data arising out of environmental applications. In this article, we review prominent approaches for the construction of non-stationary covariance functions. Once a suitable covariance function is selected, the next challenge that one faces is to carry out computation using that covariance function. Non-stationary covariance functions although flexibly capture the spatial dependence structure, model fitting with them requires $O(n^3)$ computation, which is impossible to commence if n is massive. Basis function-based construction of non-stationary covariance functions can reduce the computational cost by a large margin. Recently, the Vecchia approximation-based nearest-neighbour Gaussian process has gained popularity among applied researchers. In this article, we review these approaches and some more for the construction of scalable spatial covariance functions for point-referenced spatial data.

Key words: Geostatistical data; Non-stationary covariance function; Vecchia approximation; Scalable spatial models.

AMS Subject Classifications: 62M30.

1. Introduction

Point-referenced spatial/geostatistical data arises when observations are made at n spatial/geographical locations $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n$. They are routinely encountered in a broad range

Corresponding Author: Suman Guha
Email: suman.stat@presiuniv.ac.in

of areas like environmental, meteorological, ecological, and economic studies. The observations $y(\mathbf{s}_1), y(\mathbf{s}_2), \cdots, y(\mathbf{s}_n)$ are generally scalar-valued, and can signify temperature, rainfall, ground-level ozone concentration, house price, *etc.* Often the observations are recorded from satellite-based images and as a result, n can be as large as of the order ~ 100000. In addition to the dependent nature of the background random variables $Y(\mathbf{s}_1), Y(\mathbf{s}_2), \cdots, Y(\mathbf{s}_n)$, a large value of n renders the statistical modeling to be doubly difficult. In fact, a large n implies that the number of parameters p required for model specification may also be large. So, it is a high dimension high sample size (HDHSS) problem that couples high dimension (large p) with big data (large n).

Statistical inference is often carried out assuming that $(Y(\mathbf{s}_1), Y(\mathbf{s}_2), \cdots, Y(\mathbf{s}_n))$ is distributed according to a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. An exception to Gaussianity is noted for spatial data associated with extreme events like daily maximum windspeed, daily maximum temperature (Huser and Wadsworth (2022)), extreme snow depth (Blanchet and Davison (2011)), *etc.*, which requires modeling with multivariate extreme value distribution. Apart from that, spatial data that are positive-valued, skewed (Ayalew *et al.* (2024)) with a possible heavy tail, are also modeled by multivariate skewed distributions (Hazra *et al.* (2020)). Nevertheless, for the majority of cases, the inference boils down to estimating the mean vector and the covariance matrix of a multivariate normal distribution. Generally, the mean vector $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \mu(\mathbf{s}_2), \cdots, \mu(\mathbf{s}_n))$ is assumed to be a function of spatial locations; for example, $\mu(\mathbf{s}_i) = \beta_0 + \beta_1 \mathbf{s}_{i,1} + \beta_2 \mathbf{s}_{i,2}$. Geostatistical data $y(\mathbf{s}_1), y(\mathbf{s}_2), \cdots, y(\mathbf{s}_n)$ is often accompanied by measurements on other spatially indexed covariates, and those covariates are included in the model by extending the formula of mean to $\mu(\mathbf{s}_i) = \beta_0 + \beta_1 \mathbf{s}_{i,1} + \beta_2 \mathbf{s}_{i,2} + \gamma_1 x_1(\mathbf{s}_i) + \gamma_2 x_2(\mathbf{s}_i)$. Unlike $\boldsymbol{\mu}$, the specification of $\boldsymbol{\Sigma}$ requires additional care. A fully general specification of $\boldsymbol{\Sigma}$ as $\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ & & \vdots & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$ elicits a flexible covariance structure for the data, although it brings $\frac{n(n+1)}{2}$ parameters under consideration, thereby making statistical inference impossible. Note that, the number of data points is n, and hence, any meaningful specification of $\boldsymbol{\Sigma}$ must not exceed n parameters. One way of achieving that is to consider a parametric covariance function $c_Y(\mathbf{s}, \mathbf{s}')$ and use it to specify the spatial covariance matrix as $\boldsymbol{\Sigma} = \begin{pmatrix} c_Y(\mathbf{s}_1, \mathbf{s}_1) & c_Y(\mathbf{s}_1, \mathbf{s}_2) & \cdots & c_Y(\mathbf{s}_1, \mathbf{s}_n) \\ c_Y(\mathbf{s}_2, \mathbf{s}_1) & c_Y(\mathbf{s}_2, \mathbf{s}_2) & \cdots & c_Y(\mathbf{s}_2, \mathbf{s}_n) \\ & & \vdots & \\ c_Y(\mathbf{s}_n, \mathbf{s}_1) & c_Y(\mathbf{s}_n, \mathbf{s}_2) & \cdots & c_Y(\mathbf{s}_n, \mathbf{s}_n) \end{pmatrix}$. In that case, the estimation of $\boldsymbol{\Sigma}$ translates to the estimation of only a few unknown parameters associated with $c_Y(\mathbf{s}, \mathbf{s}')$.

A well-known and much-used parametric covariance function is the exponential covariance function defined as $c_Y(\mathbf{s}, \mathbf{s}') := \sigma^2 e^{-\phi \|\mathbf{s} - \mathbf{s}'\|_2}$. The two parameters σ^2 and $\phi > 0$ are used to specify the shape of the covariance function. σ^2 specifies the variance of the underlying spatial process $\{Y(\mathbf{s})\}$ and ϕ, which is the decay parameter, decides how strong the spatial correlation is between $Y(\mathbf{s})$ and $Y(\mathbf{s}')$. Sometimes, an additional τ^2 parameter is brought in to define a squared exponential covariance function with nugget as

$$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 & \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2 e^{-\phi\|\mathbf{s}-\mathbf{s}'\|_2} & \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2. \end{cases}$$

τ^2 is referred to as the nugget variance and it quantifies the variability of the microscale spatial components. The microscale spatial components are those parts of $\{Y(\mathbf{s})\}$ which are uncorrelated at even the minutest spatial resolution, and hence practically behave like an iid process. Besides the exponential covariance function, there are many other parametric covariance functions that are used to model geostatistical data. Some of them are presented in the following table.

Table 1: Useful parametric covariance functions for geostatistical modeling

Covariance function	Formula
Spherical	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2\left[1-\frac{3}{2}\phi\|\mathbf{s}-\mathbf{s}'\|_2+\frac{1}{2}\phi^3\|\mathbf{s}-\mathbf{s}'\|_2^3\right] \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2 \end{cases}$
Exponential	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2 e^{-\phi\|\mathbf{s}-\mathbf{s}'\|_2} \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2 \end{cases}$
Squared exponential	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2 e^{-\phi\|\mathbf{s}-\mathbf{s}'\|_2^2} \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2 \end{cases}$
Powered exponential	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2 e^{-\phi\|\mathbf{s}-\mathbf{s}'\|_2^\alpha} \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2; 0<\alpha\le 2 \end{cases}$
Rational quadratic	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \sigma^2\left(1-\frac{\|\mathbf{s}-\mathbf{s}'\|_2^2}{\phi+\|\mathbf{s}-\mathbf{s}'\|_2^2}\right) \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2 \end{cases}$
Matérn ($\nu>0$)	$c_Y(\mathbf{s},\mathbf{s}') := \begin{cases} \sigma^2+\tau^2 \text{if } 0 = \|\mathbf{s}-\mathbf{s}'\|_2 \\ \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\sqrt{2\nu}\phi\|\mathbf{s}-\mathbf{s}'\|_2\right)^\nu K_\nu\left(\sqrt{2\nu}\phi\|\mathbf{s}-\mathbf{s}'\|_2\right) \text{if } 0 < \|\mathbf{s}-\mathbf{s}'\|_2 \end{cases}$

The squared exponential covariance function resembles the exponential covariance function but suffers from the limitation that the associated spatial process $\{Y(\mathbf{s})\}$ is infinitely many times differentiable. The spherical covariance function on the other hand has compact support and hence is useful in creating a sparse $\mathbf{\Sigma}$. The Matérn covariance function is attractive in the sense that the associated spatial process $\{Y(\mathbf{s})\}$ has controllable smoothness with $\lceil\nu\rceil - 1$ times differentiability. However, all these parametric covariance functions depend only on the distance $\|\mathbf{s}-\mathbf{s}'\|_2$ disregarding the direction along which $\mathbf{s}'$ is separated from $\mathbf{s}$. This property is known as isotropy. Isotropic covariance functions are not suitable for modeling environmental datasets that are under the influence of wind flow. For such datasets, the observations separated along the direction of wind flow typically display stronger dependence compared to the ones separated along other directions. Banerjee *et al.* (2003) analyzed a scallop catch dataset where the dependence structure along different directions varied substantially thereby necessitating the use of anisotropic covariance functions.

Unlike the isotropic covariance functions, whose covariance contours are circles, the covariance contours of the anisotropic covariance functions can take the shape of arbitrary closed curves. Different notions of anisotropy have been introduced by different researchers. Zimmerman (1993) systematically studied them and classified them roughly into three different broad categories. In order to understand them we first need to define the variogram function $\gamma_Y(\mathbf{s},\mathbf{s}')$ associated with a spatial process $\{Y(\mathbf{s})\}$. It is defined by the formula

$\gamma(\mathbf{s},\mathbf{s}') := \frac{1}{2}var(Y(\mathbf{s}) - Y(\mathbf{s}'))$. In an alternative approach to geostatistical modeling, the whole theory that can be developed using the covariance function $c_Y(\mathbf{s},\mathbf{s}')$, has been developed parallel using the variogram function $\gamma(\mathbf{s},\mathbf{s}')$. Zimmerman (1993) defined three classes of anisotropy as follows. The differential dependence along the different directions is referred to as sill anisotropy if $\lim_{a\to\infty}\gamma(\frac{a\mathbf{h}}{\|\mathbf{h}\|})$ depends not on $\|\mathbf{h}\|$ but on $\mathbf{h}$. Here $\mathbf{h} = (\mathbf{s}-\mathbf{s}')$ is the lag between two spatial locations $\mathbf{s}$ and $\mathbf{s}'$. When $\lim_{a\to 0}\gamma(\frac{a\mathbf{h}}{\|\mathbf{h}\|})$ depends on $\mathbf{h}$, one says the process shows nugget anisotropy. The third and last type of anisotropy occurs when the decay parameter ϕ depends on $\mathbf{h}$. It is referred to as range anisotropy. A particularly interesting subclass of the range anisotropy is the geometric anisotropy. Geometrically anisotropic spatial process $\{Y(\mathbf{s})\}$ has a covariance function with elliptical covariance contours. A simple recipe for creating a geometrically anisotropic covariance function is to replace $\|\mathbf{s}-\mathbf{s}'\|_2$ by $\sqrt{(\mathbf{s}-\mathbf{s}')'\mathbf{A}(\mathbf{s}-\mathbf{s}')}$ in the formula of a parametric isotropic covariance function $c_Y(\mathbf{s},\mathbf{s}')$. The 2×2 matrix $\mathbf{A}$ is a pd matrix with 3 unknown parameters, that control the shape and the alignment of the elliptical covariance contours.

Although useful, the anisotropic covariance functions are not the best choice for modeling the complex dependence structure associated with geostatistical data arising out of environmental applications. The reason is that such a covariance function $c_Y(\mathbf{s},\mathbf{s}')$, although invokes differential dependence structure along different directions, is still a function of the lag $\mathbf{h}$ between two spatial locations $\mathbf{s}$ and $\mathbf{s}'$. This property is known as stationarity. Stationarity implies that the covariance between $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ remains unchanged if both the spatial locations are shifted by the same lag $\mathbf{h}$, *i.e.*, $C_Y(\mathbf{s},\mathbf{s}') = C_Y(\mathbf{s}+\mathbf{h},\mathbf{s}'+\mathbf{h})$. Efforts have been made to create non-stationary covariance functions $C_Y(\mathbf{s},\mathbf{s}')$ which depend on both $\mathbf{s}$ and $\mathbf{s}'$.

2. Towards non-stationary covariance functions

Over the years different strategies to create non-stationary covariance functions have been proposed. Here we discuss a few prominent ones.

Approach 1 : Direct construction The simplest approach is to propose a formula of $C_Y(\mathbf{s},\mathbf{s}')$ that involves both $\mathbf{s}$ and $\mathbf{s}'$ and then subsequently show that $C_Y(\mathbf{s},\mathbf{s}')$ is a valid covariance function. However, guessing such functions and then showing them to be valid covariance functions can be difficult.

Approach 2 : Transformation of the original process Alternatively, one can start with a spatial process $\{Y(\mathbf{s})\}$ that has an isotropic covariance function and then take a transformation of $\{Y(\mathbf{s})\}$ to define a new process $\{Y^*(\mathbf{s})\}$ which has an anisotropic covariance function. The transformations used are generally elementary in nature. One such transformation $Y^*(\mathbf{s}) = \sigma(\mathbf{s})Y(\mathbf{s})$ gives rise to the non-stationary covariance function of the form $C_{Y^*}(\mathbf{s},\mathbf{s}') = \sigma(\mathbf{s})\sigma(\mathbf{s}')C_Y(\mathbf{s},\mathbf{s}') = \sigma(\mathbf{s})\sigma(\mathbf{s}')f(\|\mathbf{s}-\mathbf{s}'\|_2)$. $\sigma(\mathbf{s})$ is a geographically varying positive function that enforces the departure from stationarity in a multiplicative manner. In another transformation, one can propose $Y^*(\mathbf{s}) = Y(\mathbf{s}) + \delta(\mathbf{s})Z$, where Z is a random variable with mean 0 and variance σ_Z^2 and $\delta(\mathbf{s})$ is a positive function of $\mathbf{s}$. The transformed process has the covariance function $C_Y(\mathbf{s},\mathbf{s}') + \delta(\mathbf{s})\delta(\mathbf{s}')\sigma_Z^2 = f(\|\mathbf{s}-\mathbf{s}'\|_2) + \delta(\mathbf{s})\delta(\mathbf{s}')\sigma_Z^2$. In this case, the departure from stationarity takes place in an additive manner. To combine both, one can define $Y^*(\mathbf{s}) = \sigma(\mathbf{s})Y(\mathbf{s}) + \delta(\mathbf{s})Z$ leading to a non-stationary covariance func-

tion of the form $\sigma(\mathbf{s})\sigma(\mathbf{s}')f(\|\mathbf{s}-\mathbf{s}'\|_2)+\delta(\mathbf{s})\delta(\mathbf{s}')\sigma_Z^2$. However, the class of non-stationary covariance functions that can be generated by the transformation approach are very limited.

Approach 3 : Deformation approach Richer class of non-stationary covariance functions can be created by the deformation approach. In a seminal paper, Sampson and Guttorp (1992) first came up with the idea of deformation $g(\cdot)$ of the original geographical space so that the observed spatial process $\{Y(\mathbf{s})\}$ is stationary with respect to to the deformed geographical locations $g(\mathbf{s}_1), g(\mathbf{s}_2), \cdots, g(\mathbf{s}_n)$. Hence, the covariance of $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ is of the form $f(\|g(\mathbf{s})-g(\mathbf{s}')\|_2)$. When considered in terms of the original geographical space, the covariance function is not a function of $\|\mathbf{s}-\mathbf{s}'\|_2$, hence non-stationary. This brilliant idea however suffers from the serious shortcoming that estimating the deformation function $g(\cdot)$ from the data is a highly non-linear optimization problem that can be numerically very challenging. Moreover, the estimated $g(\cdot)$ can sometimes fold over its domain leading to a meaningless covariance function, and also, the estimation process as proposed by Sampson and Guttorp (1992) requires replicated samples at the original geographical locations $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n$.

Approach 4 : Process convolution approach In time series analysis, starting with a white noise process $\{Z_t\}$ that has the simplest covariance function, one can create a moving average process $\{X_t\}$ by taking a linear combination of Z_t as $X_t := Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$. The process $\{X_t\}$ has a substantially improved covariance function compared to the original process $\{Z_t\}$. Much to the same spirit, starting with a spatial white noise process $\{Z(\mathbf{s})\}$ with a simple spatial covariance function one can create a new process $\{Y(\mathbf{s})\}$ by the following process convolution

$$Y(\mathbf{s}) := \int_{\mathbb{R}^2} K(\mathbf{s}-\mathbf{s}')Z(\mathbf{s}')d\mathbf{s}'. \tag{1}$$

Strictly speaking, the above integral is not defined and should be interpreted as $Y(\mathbf{s}) := \int_{\mathbb{R}^2} K(\mathbf{s}-\mathbf{s}')dB(\mathbf{s}')$ where $B(\mathbf{s})$ denotes a two-dimensional Brownian motion on $\mathbb{R}^2$. When interpreted as above, the process $\{Y(\mathbf{s})\}$ has a stationary covariance function given by the formula $c_Y(\mathbf{s},\mathbf{s}') = \sigma^2 \int_{\mathbb{R}^2} K(\mathbf{s}-\mathbf{t})K(\mathbf{s}'-\mathbf{t})d\mathbf{t}$. Higdon (1998) used spatially varying kernel functions in the above formula to generate a non-stationary covariance function. In that case, $Y(\mathbf{s}) := \int_{\mathbb{R}^2} K_{\mathbf{s}}(\mathbf{s}-\mathbf{s}')dB(\mathbf{s}')$, where $K_{\mathbf{s}}(\mathbf{s}-\mathbf{t})$ is a non-negative real-valued integrable function (bivariate kernel function); the associated covariance function is non-stationary, and is given by the formula

$$c_Y(\mathbf{s},\mathbf{s}') = \sigma^2 \int_{\mathbb{R}^2} K_{\mathbf{s}}(\mathbf{s}-\mathbf{t})K_{\mathbf{s}'}(\mathbf{s}'-\mathbf{t})d\mathbf{t}. \tag{2}$$

The convolution approach was later extended by Paciorek and Schervish (2006) to produce a flexible non-stationary Matérn covariance function with nugget as

$$c_{Y,NS}(\mathbf{s},\mathbf{s}') := \begin{cases} \tau^2+\sigma^2 & \text{if } \mathbf{s}=\mathbf{s}' \\ \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} |\Sigma(\mathbf{s})|^{\frac{1}{4}} |\Sigma(\mathbf{s}')|^{\frac{1}{4}} \left|\frac{\Sigma(\mathbf{s})+\Sigma(\mathbf{s}')}{2}\right|^{-\frac{1}{2}} & \\ \times \left(2\sqrt{\nu}\phi\sqrt{Q(\mathbf{s},\mathbf{s}')}\right)^{\nu} K_\nu\left(2\sqrt{\nu}\phi\sqrt{Q(\mathbf{s},\mathbf{s}')}\right) & \text{if } \mathbf{s}\neq\mathbf{s}'. \end{cases} \tag{3}$$

Here $\Sigma(\mathbf{s}) = \begin{pmatrix}\cos(\theta(\mathbf{s})) & -\sin(\theta(\mathbf{s}))\\ \sin(\theta(\mathbf{s})) & \cos(\theta(\mathbf{s}))\end{pmatrix}\begin{pmatrix}\lambda_1(\mathbf{s}) & 0\\ 0 & \lambda_2(\mathbf{s})\end{pmatrix}\begin{pmatrix}\cos(\theta(\mathbf{s})) & \sin(\theta(\mathbf{s}))\\ -\sin(\theta(\mathbf{s})) & \cos(\theta(\mathbf{s}))\end{pmatrix}$ and $Q(\mathbf{s},\mathbf{s}') = (\mathbf{s}-\mathbf{s}')'\left(\frac{\Sigma(\mathbf{s})+\Sigma(\mathbf{s}')}{2}\right)^{-1}(\mathbf{s}-\mathbf{s}')$. Paciorek and Schervish (2003) also used the $c_{Y,NS}(\mathbf{s},\mathbf{s}')$ as

the covariance function of a Gaussian process prior for a Bayesian non-parametric regression problem. While $\lambda_1(\mathbf{s})$ and $\lambda_2(\mathbf{s})$ determine the length of the major and minor axis of the elliptical covariance contours at location $\mathbf{s}$, the $\theta(\mathbf{s})$ determines the alignment of the contours. $c_{Y,NS}(\mathbf{s},\mathbf{s}')$ is non-stationary since the shape of the covariance contours vary with respect to $\mathbf{s}$.

3. Scalable covariance functions for massive geostatistical data

Once an appropriate non-stationary covariance function is selected, the next step is to estimate the unknown parameters associated with the mean and the covariance function. Assume that the vector of those unknown parameters is denoted by $\boldsymbol{\theta}$. Hence, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are functions of $\boldsymbol{\theta}$ and are better represented as $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Under the classical paradigm, the estimation is mostly carried out by maximizing the Gaussian likelihood function $L(\boldsymbol{\theta}|\mathbf{y}) := \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))}$. The likelihood function is highly non-linear in $\boldsymbol{\theta}$ and hence requires numerical algorithms for finding the global maximum. On the other hand, if the Bayesian path is chosen, one needs to find the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))}\pi(\boldsymbol{\theta})$. The posterior generally does not appear in the form of nice well-known distributions, and hence its exploration requires an MCMC method. Regardless of the classical or Bayesian approach being adopted, one needs to evaluate the terms $|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$ and $(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))$ repeatedly. Each evaluation of $|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ requires $O(n^3)$ operations and n being a large number of the order ~ 100000, the computational cost jumps to a staggering $O(10^{15})$ operations making it impossible to implement. Suitable strategies have been developed to bring the computational cost down to a manageable level. Below we discuss some such strategies. Most of these approaches are based on replacing the terms $|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$ and $(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))$ by some approximation and then carrying out the computation. Their success depends on whether the approximation to the original term is good and itself is easily computable.

3.1. Fixed rank Kriging

In one of the earliest works in this direction Cressie and Johannesson (2008), while fitting a centered Gaussian process $Y(\mathbf{s}) \sim GP(0, c_Y(\mathbf{s},\mathbf{s}'))$ to the observed geostatistical data, approximated $\{Y(\mathbf{s})\}$ by a new process defined as $\tilde{Y}(\mathbf{s}) = \sum_{r=1}^{R}\sum_{k=1}^{K_r}\theta_{r,k}\varphi_{r,k}(\mathbf{s}) + \epsilon(\mathbf{s})v(\mathbf{s})$. $\varphi_{r,k}(\mathbf{s})$ are basis functions of resolution r and $\theta_{r,k}$ are dependent Gaussian random variables with covariance matrix $K(\boldsymbol{\phi})$. So, $\tilde{Y}(\mathbf{s}) \sim GP(0, c_{\tilde{Y}}(\mathbf{s},\mathbf{s}'))$ and its covariance function $c_{\tilde{Y}}(\mathbf{s},\mathbf{s}')$ approximates $c_Y(\mathbf{s},\mathbf{s}')$. Consequently, the covariance matrix $\boldsymbol{\Sigma}$ is also being approximated by the covariance matrix $\boldsymbol{\varphi}\mathbf{K}(\boldsymbol{\phi})\boldsymbol{\varphi}' + \tau^2\mathbf{V}$. For the approximating covariance matrix $\boldsymbol{\varphi}\mathbf{K}(\boldsymbol{\phi})\boldsymbol{\varphi}' + \tau^2\mathbf{V}$ calculating the determinant and inverse it takes $O(n)$ operations only. Thus they approximated the original likelihood $L(\boldsymbol{\theta}|\mathbf{y})$ by a new likelihood $L(\boldsymbol{\phi},\tau^2|\mathbf{y})$ where $(\boldsymbol{\phi},\tau^2)$ is the vector comprising the new parameters.

3.2. Gaussian predictive process models

In another approach more geared towards the Bayesian paradigm, Banerjee *et al.* (2008) considered a centered Gaussian process $Y(\mathbf{s}) \sim GP(0, c_Y(\mathbf{s},\mathbf{s}'))$ and approximated it by a predictive process $[Y(\mathbf{s})|Y(\mathbf{s}_1^*), Y(\mathbf{s}_2^*), \cdots, Y(\mathbf{s}_k^*)] + \epsilon(\mathbf{s})$. So, the predictive process

can be expressed as $\tilde{Y}(\mathbf{s}) = [c_Y(\mathbf{s}_1, \mathbf{s}^*), c_Y(\mathbf{s}_2, \mathbf{s}^*), \cdots, c_Y(\mathbf{s}_n, \mathbf{s}^*)]{\Sigma_Y^*}^{-1} \begin{pmatrix} Y(\mathbf{s}_1^*) \\ Y(\mathbf{s}_2^*) \\ \vdots Y(\mathbf{s}_k^*) \end{pmatrix} + \epsilon(\mathbf{s}) = \mathbf{S}(\boldsymbol{\theta}) \begin{pmatrix} Y(\mathbf{s}_1^*) \\ Y(\mathbf{s}_2^*) \\ \vdots Y(\mathbf{s}_k^*) \end{pmatrix} + \epsilon(\mathbf{s})$. $\mathbf{s}_1^*, \mathbf{s}_2^*, \cdots, \mathbf{s}_k^*$ are some knot points on the geographical plane. So, the approximating covariance matrix $\Sigma_{\tilde{Y}}$ is of the form $\mathbf{S}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})\mathbf{S}'(\boldsymbol{\theta}) + \tau^2\mathbf{I}$. Calculating the determinant and inverting $\mathbf{S}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})\mathbf{S}'(\boldsymbol{\theta}) + \tau^2\mathbf{I}$ takes only $O(n)$ operations. For fixed rank Kriging, the number of basis functions $n = K_1 + K_2 + \cdots + K_R$ determines the the quality of approximation, and it should be chosen judiciously to trade quality of approximation for computational cost. In the case of the Gaussian predictive process model, the number of the knot points n plays the same role.

3.3. Covariance tapering

While the last two approaches were based on approximating the original Gaussian process $\{Y(\mathbf{s})\}$ by a new Gaussian process $\{\tilde{Y}(\mathbf{s})\}$ with which the computational cost reduces significantly to $O(n)$, other approaches directly approximate $\boldsymbol{\Sigma}$ by a new covariance matrix $\tilde{\boldsymbol{\Sigma}}$. In the covariance tapering approach, instead of approximating $\boldsymbol{\Sigma}$ by a new covariance matrix $\tilde{\boldsymbol{\Sigma}}$, one transforms $\boldsymbol{\Sigma}$ to convert it to a sparse matrix. With that, the likelihood can be rewritten as $\left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}tr((\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}))} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}tr((\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1})}$, and can be approximated by $\left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}tr((\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1})}$. Here $\boldsymbol{\Sigma}$ is approximated by $\boldsymbol{\Sigma} \odot \mathbf{T}$. The transformation $\mathbf{T}$ is referred to as the one-taper transform and it transforms $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma} \odot \mathbf{T}$. $\mathbf{T}$ is a covariance matrix formed by a compactly supported covariance function (Kaufman *et al.* (2008)). The tapered matrix $\boldsymbol{\Sigma} \odot \mathbf{T}$ is also a covariance matrix and it is sparse, thereby making the approximating likelihood scalable to massive n. A variation of the one-taper transform is called a two-taper transform that transforms $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma} \odot \mathbf{T}$ as well as the empirical covariance matrix $(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'$ to $(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})' \odot \mathbf{T}$. That way, both the model covariance matrix and the empirical covariance matrix become sparse.

3.4. Vecchia approximation and nearest neighbour Gaussian process (NNGP)

Any likelihood function can be expressed as products of conditional distributions as follows

$$[Y(\mathbf{s}_1), Y(\mathbf{s}_2), \cdots, Y(\mathbf{s}_n)] = [Y(\mathbf{s}_n) \mid Y(\mathbf{s}_{n-1}), \cdots, Y(\mathbf{s}_1)] \times [Y(\mathbf{s}_{n-1}) \mid Y(\mathbf{s}_{n-2}), \cdots, Y(\mathbf{s}_1)] \times \cdots \times [Y(\mathbf{s}_2) \mid Y(\mathbf{s}_1)] \times [Y(\mathbf{s}_1)].$$

Based on this representation Vecchia (1988) in an early work figured out how to reduce the computational cost of evaluating a likelihood. He demonstrated that the above expression is

$$\approx [Y(\mathbf{s}_n) \mid Y(\mathbf{s})_{s \in N_n}] \times [Y(\mathbf{s}_{n-1}) \mid Y(\mathbf{s})_{s \in N_{n-1}}] \times \cdots \times [Y(\mathbf{s}_2) \mid Y(\mathbf{s})_{s \in N_2}] \times [Y(\mathbf{s}_1)].$$

where N_i denotes of neighbourhood set of $\mathbf{s}_i$ that contains atmost k spatial locations. So under the traditional Gaussian setup evaluating $[Y(\mathbf{s}_i) \mid Y(\mathbf{s})_{s \in N_i}]$ requires calculating the determinant and inverse of at most a $k \times k$ covariance matrix. The computational cost

is atmost $O(k^3)$. As there are $n-1$ such products, the overall computational cost for calculating the approximating likelihood $[Y(\mathbf{s}_n) \mid Y(\mathbf{s})_{s\in N_n}] \times [Y(\mathbf{s}_{n-1}) \mid Y(\mathbf{s})_{s\in N_{n-1}}] \times \cdots \times [Y(\mathbf{s}_2) \mid Y(\mathbf{s})_{s\in N_2}] \times [Y(\mathbf{s}_1)]$ sum up to $O(nk^3)$. Although the idea was first presented by Vecchia (1988), it became familiar when Datta *et al.* (2016) applied it successfully to model a massive forest inventory dataset. Moreover, it is to the credit of Datta *et al.* (2016) who showed that the approximating likelihood can be associated with another Gaussian process, which they referred to as the nearest neighbour Gaussian process (NNGP). Although, the NNGP produced promising result, the method's dependence on the number of neighbours, and the set of neighbouring locations requires further investigation. Another issue is that the decomposition of the likelihood function as products of conditional distributions is not unique, and hence the success of the Vecchia approximation and the NNGP depend on the particular version one uses.

4. Concluding remarks

In this article, we have briefly touched upon different approaches for the creation of non-stationary covariance functions. The list is ever growing and many of them are not discussed here. For example, Fuentes (2002) considered the convolution of stationary processes and created locally stationary covariance functions. Then we have seen that the problem does not just end with the selection of an appropriate non-stationary covariance function. The advent of GIS-based data collection system coupled with advancement in data storage capacity, allows us to gather data at millions. Directly working with a non-stationary covariance function for such massive dataset leads to $O(n^3)$ computations making the task impossible to commence. In this regard, we have discussed different methods of scalable modeling of massive geostatistical data. Among them, the Vecchia approximation has recently gained popularity with the work of Datta *et al.* (2016). In a recent work, Zheng *et al.* (2023) extended the idea to a non-Gaussian spatial process. Besides the approaches discussed here, the multiresolution analysis proposed by Katzfuss (2017) is also useful in modeling massive geostatistical data. There are many more methods for scalable modeling of geostatistical data and a comparative analysis of them have been carried out in Heaton *et al.* (2019). The field is growing rapidly. For a more comprehensive review of the Bayesian methods for massive geostatistical data, one can consider the recent articles by Banerjee and Fuentes (2012) and Banerjee (2017).

Conflict of interest

The author does not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Ayalew, K. A., Manda, S., and Cai, B. (2024). Multivariate skew-normal distribution for modelling skewed spatial data. *Spatial and Spatio-temporal Epidemiology*, **51**, 100692.

Banerjee, S. (2017). High-dimensional Bayesian geostatistics. *Bayesian Analysis*, **12**, 583.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC.

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

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

Blanchet, J. and Davison, A. C. (2011). Spatial modeling of extreme snow depth. *The Annals of Applied Statistics*, **5**, 1699–1725.

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

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, **111**, 800–812.

Fuentes, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika*, **89**, 197–210.

Hazra, A., Reich, B. J., and Staicu, A.-M. (2020). A multivariate spatial skew-t process for joint modeling of extreme precipitation indexes. *Environmetrics*, **31**, e2602.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, **24**, 398–425.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, **5**, 173–190.

Huser, R. and Wadsworth, J. L. (2022). Advances in statistical modeling of spatial extremes. *Wiley Interdisciplinary Reviews: Computational Statistics*, **14**, e1537.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, **112**, 201–214.

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

Paciorek, C. and Schervish, M. (2003). Nonstationary covariance functions for gaussian process regression. *Advances in Neural Information Processing Systems*, **16**.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, **17**, 483–506.

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

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

Zheng, X., Kottas, A., and Sansó, B. (2023). Nearest-neighbor mixture models for non-gaussian spatial processes. *Bayesian Analysis*, **18**, 1191–1222.

Zimmerman, D. L. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, **25**, 453–470.