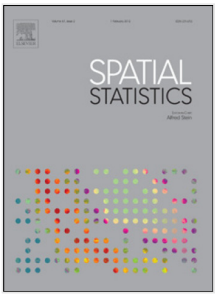
Accepted Manuscript



Spatial statistics and Gaussian preocesses:
A beautiful marriage

Alan E. Gelfand, Erin M. Schliep

PII: S2211-6753(16)30003-3
DOI: http://dx.doi.org/10.1016/j.spasta.2016.03.006
Reference: SPASTA 157

To appear in: *Spatial Statistics*

Received date: 30 December 2015
Accepted date: 30 March 2016

Please cite this article as: Gelfand, A.E., Schliep, E.M., Spatial statistics and Gaussian preocesses:
A beautiful marriage. *Spatial Statistics* (2016), http://dx.doi.org/10.1016/j.spasta.2016.03.006

Spatial Statistics and Gaussian Processes: A Beautiful Marriage

Alan E. Gelfand[1]

Duke University, Durham, North Carolina, USA

Erin M. Schliep[2]

University of Missouri, Columbia, Missouri, USA

Abstract

Spatial analysis has grown at a remarkable rate over the past two decades. Fueled by sophisticated GIS software and inexpensive and fast computation, collection of data with spatially referenced information has increased. Recognizing that such information can improve data analysis has led to an explosion of modeling and model fitting.

The contribution of this paper is to illustrate how Gaussian processes have emerged as, arguably, the most valuable tool in the toolkit for geostatistical modeling. Apart from the simplest versions, geostatistical modeling can be viewed as a hierarchical specification with Gaussian processes introduced appropriately at different levels of the specification. This naturally leads to adopting a Bayesian framework for inference and suitable Gibbs sampling/Markov chain Monte Carlo for model fitting.

Here, we review twenty years of modeling work spanning multivariate spatial analysis, gradient analysis, Bayesian nonparametric spatial ideas, directional data, extremes, data fusion, and large spatial and spatio-temporal datasets. We demonstrate that Gaussian processes are the key ingredients in all of this work. Most of the content is focused on modeling with examples being limited due to length constraints for the article. Altogether, we are able to conclude that spatial statistics and Gaussian processes do, indeed, make a beautiful marriage.

Keywords: data fusion; directional data; hierarchical model; multivariate processes; spatial Dirichlet processes; spatial gradients

1. Introduction

Spatial statistics has had an unusual history as a field within the discipline of Statistics. The stochastic process theory underlying much of the field was developed by probabilists, whereas,

[1]*Mailing address:* Department of Statistical Science, Duke University
Box 90251, Durham, North Carolina, 27708
Email address: alan@stat.duke.edu
Telephone: 919-668-5229
[2]*Mailing address:* Department of Statistics, University of Missouri
146 Middlebush Hall, Columbia, Missouri, 65211
Email address: schliepe@missouri.edu
Telephone: 573-882-4455

early on, much of the statistical methodology was developed quite independently and informally. In fact, this methodology grew primarily from the different areas of application, e.g., mining engineering, agriculture, and forestry.

As a result, for many years, spatial statistics labored on the fringe of mainstream statistics. However, the past twenty years has seen an explosion of interest in space and space-time problems. This has been largely fueled by the increased availability of inexpensive, high speed computing (as has been the case for many other areas). Such availability has enabled the collection of large spatial and spatio-temporal datasets across many fields, leading to widespread usage of sophisticated geographic information systems (GIS) software to create attractive displays along with the ability to investigate (fit and infer under) challenging, evermore appropriate and realistic models. As a result, spatial statistics has been brought into the mainstream of statistical research, changing from a somewhat ad hoc field to one that is more model-driven.

Full specification of stochastic models for the spatial process being investigated enables full inference and uncertainty assessment regarding the process. Gaussian processes (GPs) on $\mathcal{R}^2$ have become a fundamental specification in such modeling, particularly in settings where prediction is a primary goal.

Here, we focus primarily on geostatistical models, i.e., point-referenced data models. Apart from the simplest versions, such geostatistical modeling can be viewed as a hierarchical specification, with Gaussian processes introduced appropriately at different levels of the specification. Adoption of a Bayesian framework for inference and suitable Gibbs sampling/Markov chain Monte Carlo (MCMC) for model fitting follows. This is not surprising since, more generally, hierarchical modeling has emerged as the modeling paradigm for scientific work in the 21st century (e.g., Gelfand and Ghosh, 2013) and spatial statistics in particular (Gelfand, 2012).

Over the past twenty years there has been an enormous growth in such modeling. The contribution here is to highlight the substantial range of spatial settings where Gaussian processes have enabled rich and flexible specification. While our focus is on the geostatistical spatial setting, we note the importance of Gaussian processes in modeling spatial point patterns, e.g., log Gaussian Cox processes (Banerjee et al., 2014; Moller and Waagepetersen, 2003) or with lattice, grid, and areal data models using Gaussian Markov random fields (Rue and Held, 2005).

Even confining ourselves solely to the geostatistical setting, there is still too much to cover. What we offer is a review of the basic geostatistical model in hierarchical form. This leads to generalized linear spatial regression models and multivariate process models including spatially varying coefficient models. We also briefly mention simple extensions of Gaussian processes that enable more flexible process specifications. Next, we turn to elegant gradient analysis to study directionality in random realizations of spatial surfaces. Then, we consider nonparametric distributional models for spatial data. Here, rather than interpolating realizations at unobserved locations, we interpolate random distributions at unobserved locations. We next discuss spatial extremes, spatial directional data, and data fusion. We conclude with the use of Gaussian processes to accommodate large datasets.

The format of the paper is as follows. In Section 2 we clarify why Gaussian processes prove so attractive. In Section 3 we review the customary geostatistical model, viewing it hierarchically, with either a Gaussian or non-Gaussian first stage specification. In Section 4 we consider several extensions of Gaussian processes to enhance their flexibility. Section 5 considers multivariate Gaussian spatial processes while Section 6 looks at the gradient behavior associated with realizations of Gaussian processes. Section 7 turns to nonparametric extensions of Gaussian processes using Dirichlet processes. Section 8 highlights modeling of spatial extremes while Section 9 examines spatial directional data. Section 10 looks at data fusion or data assimilation

and Section 11 concludes with two Gaussian process models for large spatial datasets. We give a brief summary in Section 12.

2. Why we love Gaussian processes?

In the geostatistical setting, we need to specify random surfaces over R^2. One way to do this is through basis function representations such as splines, wavelets, radial basis functions, etc. (Ramsay and Silverman, 2002; Fahrmeir et al., 2013). These choices arise from specifying coefficients which, applied to a set of basis functions, supply surfaces. They offer explicit functions for the surfaces and are well-established in the literature. In a sense, they are not truly nonparametric since they are specified through a finite set of coefficients and, thus, provide a parametric representation of the surface.

We focus on an alternative specification where the surface is modeled as a realization of a stochastic process. Here, a realization consists of an uncountable number of random variables $\{Y(\mathbf{s}) : \mathbf{s} \in D \subset R^2\}$; it is not a function over D. We specify the *distribution* associated with such a stochastic process through finite dimensional distributions which satisfy the usual (Kolmogorov) consistency conditions. As a result, it is not surprising that Gaussian processes provide a natural specification for random surfaces. We only require a mean surface and a valid covariance function (e.g. Banerjee et al., 2014) in order to provide the multivariate normal distributions associated with any finite set of locations. Familiar distribution theory ensues in terms of marginal and conditional distributions. Computation is straightforward, particularly for Bayesian model fitting. That is, Gibbs sampling and MCMC model fitting can be implemented using standard multivariate normal random generation.

From theory, for Gaussian processes, strong stationarity is equivalent to weak stationarity. Moreover, Gaussian processes (more generally, elliptical processes (Chu, 1973; Kim and Mallick, 2003)) and suitable transformations of Gaussian processes, can support a stochastic process on $\mathcal{R}^2$.

From a modeling perspective, Gaussianity is a difficult assumption to criticize without replications. That is, given a finite set of observations from a stochastic process over a set D, we essentially see *one* realization from a multivariate distribution. How can we criticize normality with such data? Moreover, below, we customarily introduce Gaussian processes as models for spatial and/or spatio-temporal random effects. So, we are consistent with the rich literature that models random effects as normal random variables, introducing suitable variance components (West et al., 2014).

In practice, we usually focus on supplying a first order (or mean) model and a second order (or dependence) model. The first is often specified using some sort of regression and the second by selecting a suitable valid covariance function. Then, we *overlay* these choices on a Gaussian process to achieve the requisite specification.

3. Univariate point-level modeling

Consider modeling a spatial surface for temperature or for ozone levels. The basic geostatistical model for such surfaces is

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}). \quad (1)$$

The residual is partitioned into two pieces: the first is spatial ($w(\mathbf{s})$) and the second is non-spatial ($\epsilon(\mathbf{s})$). Here, $w(\mathbf{s})$ is usually a *stationary* Gaussian process with variance σ^2 while $\epsilon(\mathbf{s})$ adds the *nugget* (τ^2) effect, usually white noise (Banerjee et al., 2014; Cressie and Wikle, 2011). Interpretations attached to $\epsilon(\mathbf{s})$ are (i) pure error, (ii) measurement error, (iii) replication error or (iv) microscale variability. In any event, σ^2 and τ^2 are variance components with the inclusion of $\epsilon(\mathbf{s})$ implying that the $Y(\mathbf{s})$ surface is everywhere discontinuous.

Introducing $w(\mathbf{s})$ provides the most basic version of our Gaussian processes theme. Suppose we have data $Y(\mathbf{s}_i), i = 1, \ldots, n$, and let $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$. Under a Gaussian process for $w(\mathbf{s})$ and white noise for $\epsilon(\mathbf{s})$, we can marginalize over the $\{w(\mathbf{s}_i)\}$. With, a covariance function of the form $\mathrm{Cov}(w(\mathbf{s}), w(\mathbf{s}')) = \sigma^2\rho(\mathbf{s} - \mathbf{s}')$, we can marginalize over the w's to obtain the marginal covariance matrix, $\Sigma = \sigma^2 H(\phi) + \tau^2 I$ where $H(\phi)_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j)$. The result of such marginalization is that the covariance matrix is diagonally dominant and, therefore, well behaved with regard to inversion and determinant calculation needed for likelihood evaluation.

The resulting likelihood is given through

$$\mathbf{Y}|\boldsymbol{\theta} \sim N(X\boldsymbol{\beta}, \sigma^2 H(\phi) + \tau^2 I). \tag{2}$$

In the Bayesian setting, independent priors are usually chosen for the different parameters yielding $p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\sigma^2)p(\tau^2)p(\phi)$. Here, we are assuming $\rho(\cdot)$ to be isotropic with decay parameter ϕ. Customary candidates for priors are multivariate normal for $\boldsymbol{\beta}$ and inverse gamma for σ^2 and τ^2. Specification of $p(\phi)$ depends upon the choice of ρ; usually Uniform or Gamma priors are selected. In fact, $p(\boldsymbol{\beta})$ can even be flat (improper), but if σ^2, τ^2 and ϕ have improper priors an improper posterior results (Berger et al., 2001).

gaussain process

It is useful to recast the above specification in a hierarchical form by considering a likelihood conditional on the spatial random effects $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))$. We have

First stage:

$$\mathbf{Y}|\boldsymbol{\theta}, \mathbf{w} \sim N(X\boldsymbol{\beta} + \mathbf{w}, \tau^2 I).$$

That is, the $Y(\mathbf{s}_i)$'s are conditionally independent given the $w(\mathbf{s}_i)$'s.

Second stage:

$$\mathbf{w}|\sigma^2, \phi \sim N(\mathbf{0}, \sigma^2 H(\phi)).$$

Here we introduce a Gaussian process as a *second stage* process model.

Third stage: Add priors for $(\boldsymbol{\beta}, \tau^2, \sigma^2, \phi)$. The third stage is often referred to as a hyperprior specification.

The hierarchical model and the marginal model are equivalent. In particular, if we fit this model in a Bayesian framework, the posterior $p(\boldsymbol{\theta}|\mathbf{Y})$ is the same in either case, as is interpolation/kriging.

We have two options for model fitting. We can fit the *marginal* model, $f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, or the hierarchical model, $f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{w})p(\mathbf{w}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Generally fitting the marginal model is better. It is lower dimensional and, due to the diagonal dominance, $\sigma^2 H(\phi) + \tau^2 I$ is better *conditioned* than $\sigma^2 H(\phi)$. For the hierarchical model, conjugate full conditional distributions in σ^2 and τ^2 arise (with inverse gamma priors) and $\mathbf{w}$ has a Gaussian full conditional so MCMC updating is particularly easy. The marginal model will need Metropolis updates for σ^2, τ^2 and ϕ. In practice, however, these Metropolis updates usually work well and convergence is often faster than implementing the full Gibbs updating since we are fitting a lower dimensional version of the model. Moreover, we have not lost the $\mathbf{w}$'s under the marginalization. They are easily recovered via composition sampling (see Banerjee et al., 2014).

3.1. Spatial Generalized Linear Models

Often data sets preclude first stage Gaussian modeling. In fact, $Y(\mathbf{s})$ need not be continuous. For example, $Y(\mathbf{s})$ can be a binary or count variable. Binary examples include presence/absence of species data for observed locations $\mathbf{s}$ or presence of cloud cover at $\mathbf{s}$ with satellite data. Count data examples include abundance at a location $\mathbf{s}$ or number of accidents at a location $\mathbf{s}$. In this setting, we replace the Gaussian first stage with an exponential family member (perhaps with a dispersion parameter) but retain the second stage Gaussian process model (see Diggle et al., 1998). Now we have

First stage: $Y(\mathbf{s}_i)$'s are conditionally independent given $\boldsymbol{\beta}$ and $w(\mathbf{s}_i)$'s such that $f(y(\mathbf{s}_i)|\boldsymbol{\beta}, w(\mathbf{s}_i), \gamma)$ equals

$$h(y(\mathbf{s}_i), \gamma) \exp\left(\gamma[y(\mathbf{s}_i)\eta(\mathbf{s}_i) - \psi(\eta(\mathbf{s}_i))]\right)$$

where $g(E(Y(\mathbf{s}_i))) = \eta(\mathbf{s}_i) = \mathbf{x}^T(\mathbf{s}_i)\boldsymbol{\beta} + w(\mathbf{s}_i)$ (with $g(\cdot)$ a suitable link function) and γ is a dispersion parameter.

Second stage: Model $w(\mathbf{s})$ as a Gaussian process:

$$\mathbf{w} \sim N(\mathbf{0}, \sigma^2 H(\phi)).$$

So, continuing our theme, again we introduce a Gaussian process.

Third stage: Again, we supply hyperpriors for the model parameters.

We note that now we no longer have a stochastic process for $Y(\mathbf{s})$; rather, for a finite set of locations, we have a valid finite dimensional joint distribution. This causes no difficulty with regard to inference since we only fit the model for a finite number of points and we only interpolate at a finite number of points.

We also note that it is not sensible to add a pure error term $\epsilon(\mathbf{s})$ to the specification. These variables become redundant with the first stage conditional independence. The second stage modeling employs spatial random effects which encourages the *means* of the spatial variables at proximate locations to be similar to each other. Though we can not do explicit calculation, for locations $\mathbf{s}$ and $\mathbf{s}'$, marginal spatial dependence is induced between $Y(\mathbf{s})$ and $Y(\mathbf{s}')$.

4. Useful spatial extensions of the Gaussian process

While Gaussian processes are convenient to use and difficult to criticize, particularly as random effects specifications, they are limited in the sense that marginal distributions are always symmetric and unimodal, with fixed tail behavior. However, Gaussian processes can be employed to create extensions that enable more flexibility while still retaining computational convenience. All of these extensions can be used in geostatistical modeling.

A first example proposes extensions to t-type processes, following Palacios and Steel (2006). The simplest version specifies $W(\mathbf{s}) = \frac{Z(\mathbf{s})}{\sqrt{V/r}}$ where $Z(\mathbf{s})$ is a Gaussian process and $V \sim \chi^2_r$ is a chi-square distribution with r degrees of freedom and is independent of the process. Since the t-distribution is a scale mixture of normals, with regard to computation, we can still work with the customary multivariate normal distribution theory by merely adding a latent mixing variable in the model fitting. A richer version sets $W(\mathbf{s}) = \frac{Z(\mathbf{s})}{\sqrt{V(\mathbf{s})/r}}$. Here, the $V(\mathbf{s})$ realizations are independent of the $Z(\mathbf{s})$ surface and could be specified through a log Gaussian process or perhaps a log white noise process. In either case we sacrifice the local t-distributions but can create heavier tails and retain straightforward model fitting.

A second version is the so-called split (or double) Gaussian process (Villani and Larsson, 2006; Kottas and Krnjajić, 2009) where we define $W(\mathbf{s}) = \sigma_- Z(\mathbf{s})1(Z(\mathbf{s}) < 0) + \sigma_+ Z(\mathbf{s})1(Z(\mathbf{s}) > 0)$. Evidently, this process introduces different left and right tail behavior, hence skewness. These distributions, however, are more difficult to work with in spatial modeling because the associated n dimensional distribution arising from n spatial locations is expressed through 2^n rescalings of an n-dimensional multivariate normal.

To introduce skewness, it is more convenient to work directly with a skew Gaussian process (Zhang and El-Shaarawi, 2010). Let $W(\mathbf{s}) = \sigma\delta|Z_1(\mathbf{s})| + \sigma\sqrt{(1-\delta^2)}Z_2(\mathbf{s})$ where $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}} \in [-1, 1]$ and λ is the skewness parameter while $Z_1(\mathbf{s})$ and $Z_2(\mathbf{s})$ are independent mean 0 Gaussian processes. This is not the only construction of a skew Gaussian process in the literature (see Minozzo and Ferracuti, 2012, for a fuller discussion). However, this version is computationally convenient for spatial analysis since, given $Z_1(\mathbf{s})$, $W(\mathbf{s})$ is a Gaussian process. We can introduce the $Z_1(\mathbf{s})$'s as latent variables in model fitting.

Copulas are currently receiving increased attention as they enable flexible modeling (Nelsen, 2013). In fact, there is a substantial literature on copula processes (Jaworski et al., 2010, see, e.g.,). Focusing on Gaussian processes, in the spatial setting, suppose $Z(\mathbf{s})$ is a Gaussian process and let $W(\mathbf{s}) = G^{-1}(\Phi(Z(\mathbf{s})))$ where G is the cdf of a continuous random variable and Φ is the standard normal cdf. Then $W(\mathbf{s})$ is a stochastic process of random variables each having marginal distribution G (Sang and Gelfand, 2010). Such models are limited in their dependence structure; they inherit dependence through the inducing Gaussian process. However, computation is straightforward as they only require a component-wise monotone transformation of Gaussian variables. This allows us to, again, work with the multivariate normal distribution associated with the n locations.

Lastly, we mention the asymmetric Laplace process which was employed by Lum and Gelfand (2012) in the context of spatial quantile regression. For $p \in (0, 1)$, define $\epsilon_p(\mathbf{s}) = \sqrt{\frac{2\xi(\mathbf{s})}{\tau p(1-p)}}Z(\mathbf{s}) + \frac{1-2p}{p(1-p)}\xi(\mathbf{s})$ where $Z(\mathbf{s})$ is a Gaussian process. Let the $\xi(\mathbf{s})$ be exponential variables that can be common across all $\mathbf{s}$ or independent exponential variables at each $\mathbf{s}$. In either case, as shown in Kotz et al. (2001), marginally, $\epsilon_p(\mathbf{s})$ has an asymmetric Laplace distribution with $P(\epsilon_p(\mathbf{s}) \leq 0) = p$. This process is easy to accommodate computationally since, given $\xi(\mathbf{s})$, we again have a Gaussian process and the $\xi(\mathbf{s})$'s become latent variables in the model fitting.

5. Multivariate spatial Gaussian processes

Point-referenced spatial data often come in the form of multivariate measurements at each location. A customary example is data from environmental monitoring stations which might yield measurements on, for example, ozone, nitrous oxide, carbon monoxide, $PM_{2.5}$, and perhaps even species of $PM_{2.5}$. With such data we anticipate dependence between measurements at a particular location as well as across locations. Let $\mathbf{Y}(\mathbf{s}) : \mathbf{s} \in D$ denote a stochastic process of $p \times 1$ random vectors. We envision p spatial surfaces with dependence between the surfaces. Therefore, we need a p-dimensional stochastic process which, again, will be supplied by specifying finite dimensional distributions, e.g., for $\mathbf{Y} = (\mathbf{Y}(\mathbf{s}_1), \ldots, \mathbf{Y}(\mathbf{s}_n))$. Again, we turn to Gaussian processes but now, for n locations, we need an np-dimensional multivariate normal.

This necessitates a cross-covariance, $C(\mathbf{s}, \mathbf{s}') = Cov(\mathbf{Y}(\mathbf{s}), \mathbf{Y}(\mathbf{s}'))$, which is a $p \times p$ matrix that need not be symmetric, i.e., $cov(Y_j(\mathbf{s}), Y_{j'}(\mathbf{s}'))$ need not equal $cov(Y_{j'}(\mathbf{s}), Y_j(\mathbf{s}'))$. $C(\mathbf{s}, \mathbf{s}')$ is not positive definite except in a limiting sense; $C(\mathbf{s}, \mathbf{s})$ is the covariance matrix associated with

$\mathbf{Y}(\mathbf{s})$. With a p-variate mean specification, we will have the required Gaussian process and the associated np dimensional normal distributions

How do we create valid specification for $C(\mathbf{s}, \mathbf{s}')$? Consider the linear model of coregionalization, originally suggested for dimension reduction by Wackernagel (2003) and subsequently proposed for multivariate spatial modeling in Gelfand et al. (2004). Let $\mathbf{v}(\mathbf{s}) = A\mathbf{w}(\mathbf{s})$, $\mathbf{w}(\mathbf{s}) = (w_1(\mathbf{s}), w_2(\mathbf{s}), \ldots, w_p(\mathbf{s}))$. Here, we have p independent spatial processes, say with stationary correlation functions $\rho_j(\mathbf{s} - \mathbf{s}')$, $j = 1, 2, \ldots, p$. The cross covariance matrix (with $\mathbf{a}_j$ denoting the columns of A and $\mathbf{a}_j\mathbf{a}_j^T = T_j$) is $C(\mathbf{s}-\mathbf{s}') = \sum_{j=1}^{p} \rho_j(\mathbf{s}-\mathbf{s}')\mathbf{a}_j\mathbf{a}_j^T$. This approach is *constructive* and hence immediately provides a valid cross-covariance function. It is also stationary, and yields a distinct covariance function for each component.

A general multivariate spatial model of coregionalization using Gaussian processes emerges. That is, $\mathbf{Y}(\mathbf{s}) = \boldsymbol{\mu}(\mathbf{s}) + \mathbf{v}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s})$ with $\boldsymbol{\epsilon}(\mathbf{s}) \sim N(0, D_\epsilon)$, $(D_\epsilon)_{jj} = \tau_j^2$ and $\mathbf{v}(\mathbf{s}) = A\mathbf{w}(\mathbf{s})$ where the $w_j(\mathbf{s})$ are mean 0 Gaussian processes with individual correlation functions. Usually, $\boldsymbol{\mu}(\mathbf{s})$ arises as a regression, i.e., in the form $\mu_j(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\boldsymbol{\beta}_j$. Again, we have a hierarchical model with the first stage conditional on $\mathbf{v}(\mathbf{s})$.

An attractive example which demonstrates the power of hierarchical modeling is the spatially varying coefficient model (Gelfand et al., 2003). Suppose we specify a univariate geostatistical process $Y(\mathbf{s})$ at $\mathbf{s}$ as $Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T\boldsymbol{\beta}(\mathbf{s}) + \epsilon(\mathbf{s})$. Here, the multivariate process is for $\boldsymbol{\beta}(\mathbf{s})$ and we adopt a Gaussian linear model of coregionalization. For instance, let $p = 2$, with $\mathbf{X}(\mathbf{s})$ having a column of 1's. Then, we obtain $Y(\mathbf{s}) = \beta_0(\mathbf{s}) + X(\mathbf{s})\beta_1(\mathbf{s})$. We have a spatially varying intercept (like a spatial random effect) and a spatially varying slope. We need a bivariate spatial process because we expect the slope and intercept to be dependent at any location. We obtain a very rich class of *nonlinear* models. Perhaps what is most remarkable is that we can infer about the multivariate Gaussian process for $\boldsymbol{\beta}(\mathbf{s})$ while only observing the univariate $Y(\mathbf{s})$ process.

5.1. Kernel convolution methods

Another approach to create a multivariate Gaussian spatial process model for geostatistical data is through so-called moving averages (Barry and Ver Hoef, 1996), also referred to as kernel convolution (Higdon et al., 2002). In fact, this is a special case of a more general approach to create stationary processes (Yaglom, 1987, Chapter 8). Suppose $k_l(\cdot), l = 1, ..., p$ is a set of p square integrable kernel functions on $\Re^2$ with $k_l(\mathbf{0}) = 1$. Let $w(\mathbf{s})$ be a mean 0, variance 1 Gaussian process with correlation function ρ. Define the p-variate spatial process $\mathbf{v}(\mathbf{s})$ by $v_l(\mathbf{s}) = \sigma_l \int k_l(\mathbf{s}-\mathbf{t})w(\mathbf{t})d\mathbf{t}$, $l = 1, ..., p$. Then, $\mathbf{v}(\mathbf{s})$ is a mean 0 multivariate Gaussian process with cross-covariance function $C(\mathbf{s}, \mathbf{s}')$ having (l, l') entry

$$(C(\mathbf{s}, \mathbf{s}'))_{ll'} = \sigma_l\sigma_{l'} \int\int k_l(\mathbf{s} - \mathbf{t})k_{l'}(\mathbf{s}' - \mathbf{t}')\rho(\mathbf{t} - \mathbf{t}')d\mathbf{t}d\mathbf{t}'. \quad (3)$$

By construction, $C(\mathbf{s}, \mathbf{s}')$ is valid. By change-of-variable, $(C(\mathbf{s}, \mathbf{s}'))_{ll'}$ depends only on $\mathbf{s} - \mathbf{s}'$, i.e., $\mathbf{v}(\mathbf{s})$ is a stationary process. If the k_l depend upon $\mathbf{s}-\mathbf{s}'$ only through $\|\mathbf{s}-\mathbf{s}'\|$ and ρ is isotropic, then $C(\mathbf{s} - \mathbf{s}')$ is isotropic.

In practice, working with this process requires discrete approximation, employing a finite set of *knot* locations. Specifically, we choose a finite set of locations $\mathbf{t_1}, ...\mathbf{t_r}$ and define $v_l(\mathbf{s}) = \sigma_l \sum_{j=1}^{r} k_l(\mathbf{s} - \mathbf{t}_j)w(\mathbf{t}_j)$. Now, $(C(\mathbf{s}, \mathbf{s}'))_{ll'}$ is a double sum but the resulting Gaussian process is no longer stationary.

6. Spatial gradients

Often it is of interest to study the gradient behavior of a spatial surface. For example, with a surface of land values, where is it relatively flat, where is it rising rapidly, or, at a given point, in what direction it is changing most rapidly. For a parametric function, we do this using usual calculus methods. In a regression setting, we would refer to this as a sensitivity analysis.

When the surface is a realization of a stochastic process, we need different methods. To calculate a derivative at a given location on a random surface we need some notion of smoothness for the surface. For Gaussian processes with mean square differentiable realizations, we can develop some elegant theory (Banerjee et al., 2003; Banerjee and Gelfand, 2003).

Define $Y(\mathbf{s})$ to be mean square differentiable at $\mathbf{s}_0$ if there exists a vector $\nabla_Y(\mathbf{s}_0)$ such that, for any scalar h and unit vector $\mathbf{u}$, $Y(\mathbf{s}_0 + h\mathbf{u}) = Y(\mathbf{s}_0) + h\mathbf{u}^T\nabla_Y(\mathbf{s}_0) + r(\mathbf{s}_0, h\mathbf{u})$ (we say that $Y(\mathbf{s})$ is first order linear) where $r(\mathbf{s}_0, h\mathbf{u}) \to 0$ in the L_2 sense as $h \to 0$. That is, for any unit vector $\mathbf{u}$,

$$\lim_{h\to 0} E\left(\frac{Y(\mathbf{s}_0 + h\mathbf{u}) - Y(\mathbf{s}_0) - h\mathbf{u}^T\nabla_Y(\mathbf{s}_0)}{h}\right)^2 = 0. \tag{4}$$

We note that first order linearity is needed to ensure that a mean square differentiable process is mean square continuous.

With $\mathbf{u}$ a unit vector, let $Y_{\mathbf{u},h}(\mathbf{s}) \equiv \frac{Y(\mathbf{s}+h\mathbf{u})-Y(\mathbf{s})}{h}$ be the finite difference at $\mathbf{s}$ in direction $\mathbf{u}$ at scale h. For a fixed $\mathbf{u}$ and h, $Y_{\mathbf{u},h}(\mathbf{s})$ is a well-defined process on R^d; it is the finite difference process at scale h in direction $\mathbf{u}$. If $Y(\mathbf{s})$ is a Gaussian process, it is clear that $Y_{\mathbf{u},h}(\mathbf{s})$ is a Gaussian process whose covariance function can be easily calculated.

Now, let $D_{\mathbf{u}}Y(\mathbf{s}) = \lim_{h\to 0} Y_{\mathbf{u},h}(\mathbf{s})$, assuming the limit exists. If $Y(\mathbf{s})$ is mean square differentiable in R^d, i.e., the limit in (4) holds for every $\mathbf{s}_0$ in R^d, then for each $\mathbf{u}$,

$$D_{\mathbf{u}}Y(\mathbf{s}) = \lim_{h\to 0}\frac{Y(\mathbf{s}+h\mathbf{u}) - Y(\mathbf{s})}{h} = \lim_{h\to 0}\frac{h\mathbf{u}^T\nabla_Y(\mathbf{s}) + r(\mathbf{s}, h\mathbf{u})}{h} = \mathbf{u}^T\nabla_Y(\mathbf{s}).$$

$D_{\mathbf{u}}Y(\mathbf{s})$ is a well-defined process on R^d which we call the directional derivative process in the direction $\mathbf{u}$.

These directional derivative processes have several attractive properties. Again, if $Y(\mathbf{s})$ is a Gaussian process, these directional derivative processes are also Gaussian processes. If the unit vectors $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_d$ form an orthonormal basis set for R^d, $\mathbf{u}$ in R^d can be written as $\mathbf{u} = \sum_{i=1}^d w_i\mathbf{e}_i$ with $w_i = \mathbf{u}^T\mathbf{e}_i$ and $\sum_{i=1}^d w_i^2 = 1$. Then

$$D_{\mathbf{u}}Y(\mathbf{s}) = \mathbf{u}^T\nabla_Y(\mathbf{s}) = \sum_{i=1}^d w_i\mathbf{e}_i^T\nabla_Y(\mathbf{s}) = \sum_{i=1}^d w_i D_{\mathbf{e}_i}Y(\mathbf{s}). \tag{5}$$

To study directional derivative processes in arbitrary directions, we need only work with a basis set of directional derivative processes; in R^2, we only need the East and the North, directions, (1,0) and (0,1), respectively.

By the Cauchy-Schwarz inequality, at any location, we immediately have the direction of the maximum gradient, $\nabla_Y(\mathbf{s})$, and the magnitude of the gradient in this direction, $\|\nabla_Y(\mathbf{s})\|$.

If we start with a stationary Gaussian process whose covariance function has continuous second order (and mixed derivatives), then the directional derivative processes are also stationary and, using chain rule calculations, we can obtain the associated covariance matrix (see Banerjee et al., 2003, for details). Lastly, we can consider the trivariate Gaussian process $(Y(\mathbf{s}), D_{(1,0)}Y(\mathbf{s}), D_{(0,1)}Y(\mathbf{s}))^T$ and obtain the associated cross-covariance function (again, see Banerjee et al., 2003).

7. Spatial Dirichlet Processes

In some geostatistical settings, normality and convenient stationarity usually adopted with a Gaussian process may not be rich enough. In Section 4 we suggested extensions of Gaussian processes to enrich behavior. We now take up a much different enrichment; a nonparametric spatial modeling approach using the Dirichlet process (DP) which employs Gaussian process realizations as atoms (Gelfand et al., 2005). Applied to a finite set of locations, the resulting joint distribution is *random*. Though induced through Gaussian processes, these distributions are non-Gaussian, nonstationary, and, in fact, are a.s. discrete. We use kernel mixing (so-called mixtures of Dirichlet processes (MDPs)) to provide random continuous distributions. For the process $w(\mathbf{s})$, we have random $\{F(w(\mathbf{s})) : \mathbf{s} \in D\}$. We create the $F(w(\mathbf{s}))$ to be dependent and such that the realized $F(w(\mathbf{s}))$ converges to the realized $F(w(\mathbf{s}_0))$ as $\mathbf{s} \to \mathbf{s}_0$. What we achieve is a form of kriging but we see that it is spatial prediction for distributions.

The DP creates a random distribution by assigning random weights to random atoms (Rodriguez and Kottas, 2014). The weights, $\{p_j, j = 1, 2, ...\}, \sum_j p_j = 1$, come from stickbreaking (Rodriguez and Kottas, 2014). The atoms come from a base *distribution*. The random distribution is equivalent to a countably infinite set of random probabilities summing to 1 a.s. along with an associated countably infinite set of atoms. For us, each atom is a realization of a Gaussian process over D, say $\mathbf{w}^*_{D,j}$. So, "$G \Leftrightarrow (\{p_j\}, \{\mathbf{w}^*_{D,j}\})$". The geostatistical model randomly chooses spatial surfaces from a countably infinite set of Gaussian process realizations. Therefore, the DP realization is a.s. discrete; at any location $\mathbf{s}$, $w(\mathbf{s})$ takes on only a countably infinite set of values. When we fit the model, we will have a set of, say, n spatial samples, where each sample is observed at a finite set of spatial locations. The DP assigns a label, i.e., a choice of atom, to each sample. So, the DP allows for ties meaning that, for example, with random probability p_j^2, two samples will be assigned the same atom, $\mathbf{w}^*_{D,j}$, i.e., the same Gaussian process realization.

DP mixing overcomes the a.s. discreteness of G, i.e., if $\mathbf{w}_D$ given G is a realization from G and $\mathbf{Y}_D - \mathbf{w}_D$ is a realization from a pure error process, then, DP mixing is a convolution; it smooths *discreteness* in G. Formally, $F\left(\mathbf{Y}_D \mid G, \tau^2\right) = \int \mathcal{K}\left(\mathbf{Y}_D - \mathbf{w}_D \mid \tau^2\right) G\left(d\mathbf{w}_D\right)$ and formally, *differentiating* to densities, $f\left(\mathbf{Y}_D \mid G, \tau^2\right) = \int k\left(\mathbf{Y}_D - \mathbf{w}_D \mid \tau^2\right) G\left(d\mathbf{w}_D\right)$. As a result, $Y(\mathbf{s}) = w(\mathbf{s}) + \epsilon(\mathbf{s})$ where $w(\mathbf{s})$ arises under the above spatial DP prior model and $\epsilon(\mathbf{s})$ is $\mathrm{N}(0, \tau^2)$, a pure error (nugget) component. Apart from the spatial mean, we have the familiar spatial and pure error terms as in (1). However, we see that we are convolving distributions rather than convolving process variables to create the $Y(\mathbf{s})$ process.

The finite dimensional joint density of $\mathbf{Y} = (Y(\mathbf{s}_1), ..., Y(\mathbf{s}_m))'$, given $G^{(m)}$, where $G^{(m)} \sim DP(\nu G_0^{(m)})$, and given τ^2, is $f\left(\mathbf{Y} \mid G^{(m)}, \tau^2\right) = \int N_n\left(\mathbf{Y} \mid \mathbf{w}, \tau^2 I_n\right) G^{(m)}\left(d\mathbf{w}\right)$. The a.s. representation of $G^{(m)}$ yields that $f(\mathbf{Y} \mid G^{(m)}, \tau^2)$ is a.s. of the form $\sum_{l=1}^{\infty} p_l N_n(\mathbf{Y} \mid \mathbf{w}^*_l, \tau^2 I_n)$, i.e., a countable location mixture of normals.

Usually we add a regression term $X^T\boldsymbol{\beta}$, to the kernel of the mixture model. As a result, with n replicates, the following semiparametric hierarchical model emerges:

$$
\begin{array}{ccc}
\mathbf{Y}_i \mid \mathbf{w}_i, \boldsymbol{\beta}, \tau^2 & \overset{ind.}{\sim} & N_n(\mathbf{Y}_i \mid X_i^T\boldsymbol{\beta} + \mathbf{w}_i, \tau^2 I_n) \\
\mathbf{w}_i \mid G^{(m)} & \overset{i.i.d.}{\sim} & G^{(m)}, i = 1, ..., n \\
G^{(m)} \mid \nu, \sigma^2, \phi & \sim & DP(\nu G_0^{(m)}) \\
\boldsymbol{\beta}, \tau^2 & \sim & N_p(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \Sigma_\beta)\mathrm{IG}(\tau^2 \mid a_\tau, b_\tau) \\
\nu, \sigma^2, \phi & \sim & \mathrm{G}(\nu \mid a_\nu, b_\nu)\mathrm{IG}(\sigma^2 \mid a_\sigma, b_\sigma)[\phi].
\end{array}
$$

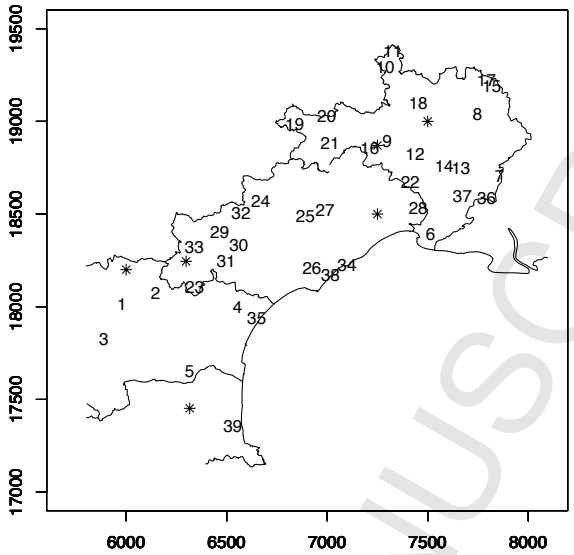


Figure 1: Map of the Languedoc-Roussillon region in southern France showing the 39 sites where the precipitation data have been observed. The six hold-out sites considered for spatial prediction in the simulation experiment are denoted by $*$. The boundaries of three French departments are also drawn.

So, $G_0^{(m)}(\cdot \mid \sigma^2, \phi) = N_n(\cdot \mid 0_n, \sigma^2 H_n(\phi))$. Model fitting is standard and kriging can also be implemented. See Gelfand et al. (2005) for details.

We present a brief illustration showing the potential benefit of working with a spatial DP model using precipitation data from the Languedoc-Rousillon Region in Southern France. These data were first discussed in Meiring et al. (1997) and again in Damian et al. (2001). We consider 75 altitude-adjusted 10-day aggregated precipitation records for the 39 sites in Figure 1. A version of the model above was fitted (see Gelfand et al., 2005, for full details) to 33 sites, removing two (denoted by $*$) in each of three subregions, in the interest of validation.

To examine nonstationarity, 21 pairs of sites were selected across the region, all pairs having the same separation vector. The pairs are shown in the top panel of Figure 2. The bottom panel shows the associated posterior mean covariances under the model fitting, suggesting departure from stationarity in support of a nonstationary spatial specification.

Figure 3 provides results for the held out sites $\mathbf{s}_4, \mathbf{s}_{35}, \mathbf{s}_{29}, \mathbf{s}_{30}, \mathbf{s}_{13}, \mathbf{s}_{37}$. Posterior predictive densities are compared with the data at these sites. We see that the spatial DP model captures the held out data well.

7.1. Generalized Spatial DPs

The spatial DP is a very strict labeling process. It provides global surface selection; i.e., for a particular sample, the same label is assigned to every location. It would be attractive to allow local surface selection, that is, for different $\mathbf{s}$, to have a spatial DP that allows a different choice of atom. Arguably, we might require fewer random atoms/surfaces to explain our observations with local selection than by insisting on the same label selection for all $\mathbf{s}$.

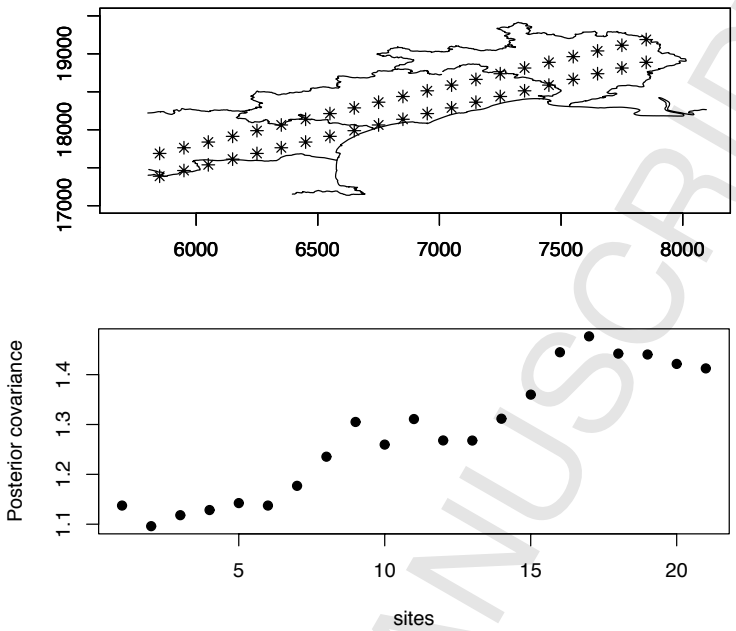


Figure 2: French precipitation data: investigating stationarity. (top) Twenty-one pairs of sites with all pairs having the same separation vector. (bottom) Associated posterior mean covariances.

The issue here then, is not specification of the Gaussian process for the atoms. The challenge is the specification of a labeling process, i.e., a specification that provides, for any set $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m\}$, a vector of random labels, $(L(\mathbf{s}_1), L(\mathbf{s}_2), ..., L(\mathbf{s}_m))$. We need to provide such selection for any number and choice of locations. We want spatial structure in selection in the sense that the closer two locations are the more likely they are to select the same surface. We might think in terms of *hybrid* curves or *canonical* curves. As an example, in brain imaging studying neurological activity level, we consider healthy brain images (surfaces) as well as impaired brain images. Typically, only a portion of the brain is impaired suggesting surface selection according to where the brain is damaged and where it is healthy.

For m locations, we need to specify labels to provide $pr(\theta(\mathbf{s}_1) = \theta^*_{l_1}(\mathbf{s}_1), \theta(\mathbf{s}_2) = \theta^*_{l_2}(\mathbf{s}_2)), ..., \theta(\mathbf{s}_m) = \theta^*_{l_m}(\mathbf{s}_m))$. Several constructions have been considered in the literature. They include multivariate stickbreaking using a countable collection of Gaussian process realizations (Duan et al., 2007) and a copula transformation of a Gaussian process to a k-dimensional hypercube (Petrone et al., 2009). The objective is to specify a random joint labeling distribution such that the probability is high that $L(\mathbf{s}) = L(\mathbf{s}')$ when $\|\mathbf{s} - \mathbf{s}'\|$ is small.

8. Modeling spatial extremes

There is often interest in studying extreme events, particularly with regard to weather and environmental exposure. For instance, we can imagine a collection of monitoring stations recording such data over years and we might be interested in the geostatistical problem of spatially modeling annual maxima. We envision spatial structure in the maxima; locations close to each other

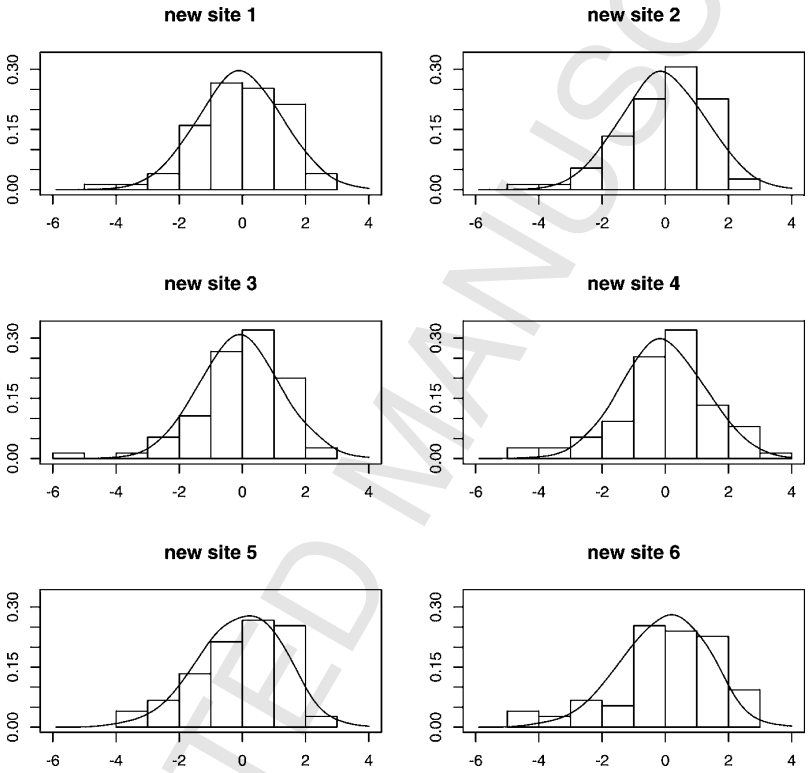


Figure 3: French precipitation data: Posterior predictive densities at new sites ($\mathbf{s}_4, \mathbf{s}_{35}, \mathbf{s}_{29}, \mathbf{s}_{30}, \mathbf{s}_{13}$, and $\mathbf{s}_{37}$), based on the model fitted to data after removing these six sites. The histograms are based on the data observed at the corresponding held out sites.

will likely exhibit similar maxima. The distributions of the extremes are not Gaussian but we can use Gaussian processes to accommodate modeling for this setting.

Focusing on the maximum, suppose we have a sequence $Y_1, Y_2, ...$ of iid random variables and, for a given n, we seek a parametric model for $M_n = \max(Y_1, ..., Y_n)$. If the distribution of the Y_i is specified, the exact distribution of M_n is known. Otherwise, we usually turn to extreme value theory which considers the existence of $\lim_{n\to\infty} Pr((M_n - b_n)/a_n \leq y) \equiv F(y)$ for two sequences of real numbers $a_n > 0, b_n$. If $F(y)$ is a non-degenerate distribution function, only three limiting types exist: the Gumbel, the Fréchet, or the Weibull. All can be usefully expressed under the umbrella of the generalized extreme value distribution (GEV) (Coles et al., 2001).

The GEV distribution is:

$$G(y; \mu, \sigma, \xi) = \exp\left\{-\left[1 + \xi\left(\frac{y-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{6}$$

for $\{y : 1 + \xi(y - \mu)/\sigma > 0\}$. Here, $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ is the scale parameter and $\xi \in \mathbb{R}$ is the shape parameter. The *residual* $V = (Y - \mu)/\sigma$ follows a GEV$(0, 1, \xi)$. If we let $Z = (1 + \xi V)^{\frac{1}{\xi}}$, equivalently, $V = \frac{Z^{\frac{1}{\xi}}-1}{\xi}$, then Z follows a standard Fréchet distribution, with distribution function $\exp(-z^{-1})$ (perhaps more familiar as an inverse gamma, $IG(1, 1)$).

We can specify a continuous spatial extremes process model hierarchically. In particular, at the first stage we have:

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \frac{\sigma(\mathbf{s})}{\xi(\mathbf{s})}(Z(\mathbf{s})^{\xi(\mathbf{s})} - 1) \tag{7}$$

where $Z(\mathbf{s})$ follows a standard Fréchet distribution.

The second stage would provide a *process* model for $\mu(\mathbf{s})$ (and possibly for $\sigma(\mathbf{s})$ and/or $\xi(\mathbf{s})$, although these are more challenging to identify and to learn about). In the spirit of this paper, we would adopt a Gaussian process specification for $\mu(\mathbf{s})$ (Sang and Gelfand, 2009). We view $Z(\mathbf{s})$ as the "standardized residual" in the first stage GEV model. Following the framework of the geostatistical model (1), suppose we assume the $Z(\mathbf{s})$ are conditionally independent, i.e., a "pure error" first stage assumption. Then, the $Z(\mathbf{s})$ are *i.i.d*. Note that even if the surface for each model parameter is smooth, the realized extremes surface will be everywhere discontinuous.

To smooth the extremes surface, recalling the copula discussion in Section 4, we can transform a standard spatial Gaussian process with mean 0, variance 1, and correlation function $\rho(\mathbf{s}, \mathbf{s}'; \theta)$, denoted by $W(\mathbf{s})$, to a spatial extreme value process. More precisely, we define the *standard* Fréchet spatial process as the transformed Gaussian process created by $Z(\mathbf{s}) = G^{-1}(\Phi(W(\mathbf{s}))$, G the cdf of a standard Fréchet distribution. $Z(\mathbf{s})$ is a valid stochastic process since it is induced by a strictly monotone transformation of a Gaussian process. As such, the dependence structure in $Z(\mathbf{s})$ is completely determined by $\rho(\mathbf{s}, \mathbf{s}'; \theta)$. With observed extremes at sites $\{s_i, i = 1, 2, ..., n\}$, we start with the joint multivariate normal distribution of $\mathbf{W} = (W(\mathbf{s}_1), ..., W(\mathbf{s}_n))$. This distribution, using the copula transformation, induces the joint distribution for $\mathbf{Z} = (Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n))$.

We note a potential reservation regarding the use of the Gaussian process here. The joint tail dependence behavior of $Z(\mathbf{s})$ and $Z(\mathbf{s}')$ is not the same as we would create from modeling the joint distribution using max-stable processes which are customary in extreme value theory (Coles et al., 2001). With max-stable processes, the tail behavior is controlled by the extremal coefficient, $\frac{\log P(Z(\mathbf{s})\leq z, Z(\mathbf{s}')\leq z)}{\log P(Z(\mathbf{s})\leq z)} \in [1, 2]$. With the Gaussian process, however, this coefficient is 2, regardless of $\mathbf{s}$ and $\mathbf{s}'$; we can only create tail independence. Unfortunately, joint distributions for max-stable processes can not be written down explicitly for more than three dimensions so

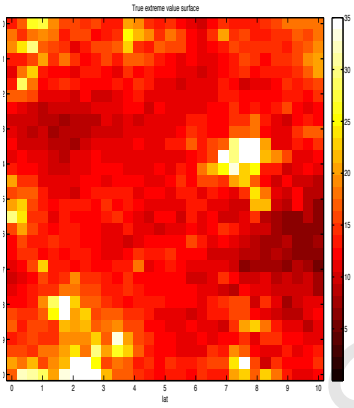


Figure 4: Observed rainfall surface for 2006.

we can not provide explicit likelihoods for a large collection of spatial locations. The current solution has been to employ a pseudo-likelihood (Padoan et al., 2010).

Returning to the Gaussian copula specification, again we have a hierarchical model and we might adopt a Gaussian process specification for $\mu(\mathbf{s})$. That is, suppose $\mu(\mathbf{s}) = X(\mathbf{s})^T\beta + V(\mathbf{s})$ where $X(\mathbf{s})$ is a site-specific vector of covariates. The $V(\mathbf{s})$ are spatial random effects specified as a mean 0 Gaussian process. Plugging in the model for $\mu(\mathbf{s})$, we obtain

$$Y(\mathbf{s}) = X^T(\mathbf{s})\beta + V(\mathbf{s}) + \frac{\sigma}{\xi}(Z(\mathbf{s})^{\xi} - 1) \tag{8}$$

with $Z(\mathbf{s}) = G^{-1}\Phi(W(\mathbf{s}))$. We can contrast this with (7). Here, we have two sources of spatial dependence, $Z(\mathbf{s})$ and $V(\mathbf{s})$, each with an associated range of dependence. That is, we have multiscale spatial dependence, both handled through Gaussian processes. Naturally, and for identifiability, we assign the shorter range spatial dependence to the $Z(\mathbf{s})$ process since it is introduced only to provide local smoothing. See Sang and Gelfand (2010) for details on fitting such models.

We provide a brief illustration using rainfall data from the Cape Floristic Region in South Africa as discussed in Sang and Gelfand (2010). The data consist of daily precipitation surfaces arising via interpolation to 1332 grid cells at 10 km resolution based on records reported by up to 3000 stations over South Africa over the period from 1950-2006. We consider derived annual maxima surfaces from the study period as our *observed* data. In Figure 4 we show the observed gridded surface for 2006. In Figure 5 we show (left) the fitted surface using the model in (7) and (right) the fitted surface using the copula smoothing from (8). Clearly, the smoothed version better fits the observed data. See Sang and Gelfand (2010) for fuller discussion.

9. Directional data

Directional data, also referred to as circular or angular data, consider measurements that are on the circle. Since this type of data are less prevalent than *linear* data, the field is less developed from a modeling and inferential perspective. Applications include meteorology, e.g., measurements of wind directions, oceanography, e.g., measurements of wave directions, ecology, e.g., animal movement directions, as well as data that are periodic in time, say daily. In this last

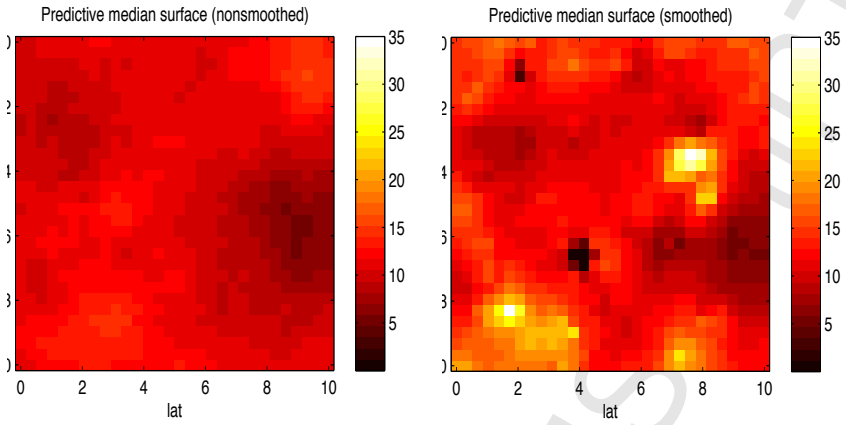


Figure 5: (left) The predictive median surface for extreme values using model (7) with non-smoothed residuals. (right) The predictive median surface for extreme values using model (8) with smoothed residuals.

context, we are thinking of clock time. That is, we are "wrapping" it to be circular in order to record, say, the daily time of maximum ozone level, or the time of day of a particular type of crime; we convert $[0:00-24:00)$ to $[0, 2\pi)$ by scaling.

Spatial applications include wind or wave directions at spatial locations. Collecting such data over time leads to spatio-temporal considerations. Moreover, we can observe associated linear variables, e.g., pollutant levels with wind direction, wave heights with wave directions. Challenges include specification of the joint spatial models and regressions - whether they should be circular on linear or vice versa.

To start, we need to specify a circular distribution, i.e., a distribution whose entire mass is on the circumference of a unit circle. We work with the absolutely continuous case (w.r.t. Lebesgue measure on the circle) and denote the random variable by θ with density $f(\theta) \geq 0$. We require the following properties: (i) for the cdf, $F(\theta + 2\pi) - F(\theta) = 1$, $\theta \in R^1$; (ii) $\int_0^{2\pi} f(\theta)d\theta = 1$; and (iii) $f(\theta + 2k\pi) = f(\theta)$ for any integer k (so that f is periodic).

An important challenge is that the circle is topologically different from the line. Angular data has *no* magnitude; a value is only assigned given an orientation but inference should not depend on the choice of origin. Moreover, without magnitudes, it requires care to think about correlation (see Jammalamadaka and Sengupta (2001)).

We consider two approaches for specifying circular distribution models. Both are built through normal distributions, hence both can be extended to circular geostatistical settings by employing Gaussian processes. First, suppose we wrap a linear variable, i.e., let $\theta = Y \mathrm{mod} 2\pi$. If $g(y)$ is a density on R^1, the associated wrapped density looks like

$$f(\theta) = \sum_{k=-\infty}^{\infty} g(\theta + 2\pi k) \tag{9}$$

For any circle of circumference L, we can rescale from $[0, L)$ to $[0, 2\pi)$.

We obtain a multivariate version with a multivariate density g on R^p,

$$f(\boldsymbol{\theta}) = \sum_{k_1=-\infty}^{\infty} ... \sum_{k_p=-\infty}^{\infty} g(\boldsymbol{\theta} + 2\pi\mathbf{k}). \tag{10}$$

Evidently, a convenient choice for g in (10) is a multivariate normal. In (10), $\boldsymbol{\theta}$ is the observed vector of angles, $\boldsymbol{\theta} + 2\pi\mathbf{K}$ is the linear vector, and $\mathbf{K}$ is the latent vector of *winding* numbers, the number of times each of the components of the latent linear variable was wound to put the components on $[0, 2\pi)$. The joint density for $(\boldsymbol{\theta}, \mathbf{K})$ is $f(\boldsymbol{\theta}, \mathbf{k}) = g(\boldsymbol{\theta} + 2\pi\mathbf{k})$.

Since Gaussian processes are specified through their finite dimensional distributions, we can immediately induce a wrapped Gaussian process from a linear Gaussian process. If a linear Gaussian process has covariance function $\sigma^2\rho(\mathbf{s}-\mathbf{s}';\phi)$, then we can write $\boldsymbol{\theta} = (\theta(\mathbf{s}_1), \theta(\mathbf{s}_2), ..., \theta(\mathbf{s}_n)) \sim WN(\mu\mathbf{1}, \sigma^2 R(\phi))$ where $R(\phi)_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j;\phi)$. Evidently, the wrapped Gaussian process inherits the association structure of the linear process. Model fitting, inference, and kriging for a wrapped Gaussian process is straightforward, introducing the latent winding numbers into the Gibbs sampling (see Jona Lasinio et al., 2012). Since the marginal distributions under the Gaussian process are symmetric and unimodal, so will be the marginals under the wrapped Gaussian process. However, we can envision wrapping the extensions of the Gaussian process discussed in Section 4. For example, we can wrap the skew Gaussian process (Zhang and El-Shaarawi, 2010). See Mastrantonio et al. (2015) for a full development.

As a second model, we consider a projection approach, i.e., embedding to the unit circle within R^2. Suppose $\mathbf{U} = (U_1, U_2) \sim g(u_1, u_2)$, a density on R^2. Then, $\mathbf{V} = (V_1, V_2) = (\frac{U_1}{\|\mathbf{U}\|}, \frac{U_2}{\|\mathbf{U}\|})$, where $\|\mathbf{U}\|$ is the length of $\mathbf{U}$, making $\mathbf{V}$ a point on the unit circle. The associated angle is $\theta = \arctan^*\frac{V_2}{V_1} = \arctan^*\frac{U_2}{U_1}$[3]. In fact, we have the usual polar coordinate relationships, $U_1 = R\cos\theta$ and $U_2 = R\sin\theta$ with $R = \|\mathbf{U}\|$. Here, θ is observed but R is latent.

This suggests the projected normal (PN) distribution as a model for angular data, i.e., if $\mathbf{U} \sim N_2(\boldsymbol{\mu}, \Sigma)$, then $\theta \sim PN_2(\boldsymbol{\mu}, \Sigma)$. The projected normal distribution is more flexible than the wrapped normal; it can be asymmetric and bimodal. In fact, we can offer a nice characterization. The collection of mixtures of projected normals is dense in the class of all circular distributions. See Wang and Gelfand (2013) for fuller discussion of the above. The density function for θ is very messy. However, if we introduce R as a latent variable, then the joint density for (θ, R) is routine to obtain, as a polar coordinate transformation, from the bivariate normal density for $\mathbf{U}$. So, inference using Gibbs sampling is straightforward for the projected normal distribution.

Hence, we can specify the projected Gaussian process. For $\{\theta(\mathbf{s}_i), i = 1, 2, ..., n\}$, we assume the latent $\mathbf{U}(\mathbf{s}_i)$ come from a bivariate Gaussian process. This induces a spatial process for the $\theta(\mathbf{s}_i)$ which we call the projected normal Gaussian process. From Section 5, there are many ways to specify a bivariate spatial Gaussian process. Since we are only interested in inducing a univariate process, a separable cross-covariance function, $C(\mathbf{s}, \mathbf{s}') = \rho(\mathbf{s}, \mathbf{s}')T$, is all we need. Again, model fitting, inference and kriging are all straightforward with the introduction of latent R's. Specifically, we can krige posterior predictive samples of say, $\mathbf{U}(\mathbf{s}_0)$, which in turn induce posterior predictive samples of $\theta(\mathbf{s}_0)$ (Wang and Gelfand, 2014).

The paper of Wang and Gelfand (2014) presents a spatial analysis of wave direction data in the Adriatic Ocean of the coast of Italy. Here, in the interest of saving space, we simply show

[3] From Jammalamadaka and Sengupta (2001), $\arctan^*(S/C)$ is formally defined as $\arctan(S/C)$ if $C > 0, S \geq 0$; $\pi/2$ if $C = 0, S > 0$; $\arctan(S/C) + \pi$ if $C < 0$; $\arctan(S/C) + 2\pi$ if $C \geq 0, S < 0$; undefined if $C = 0, S = 0$.

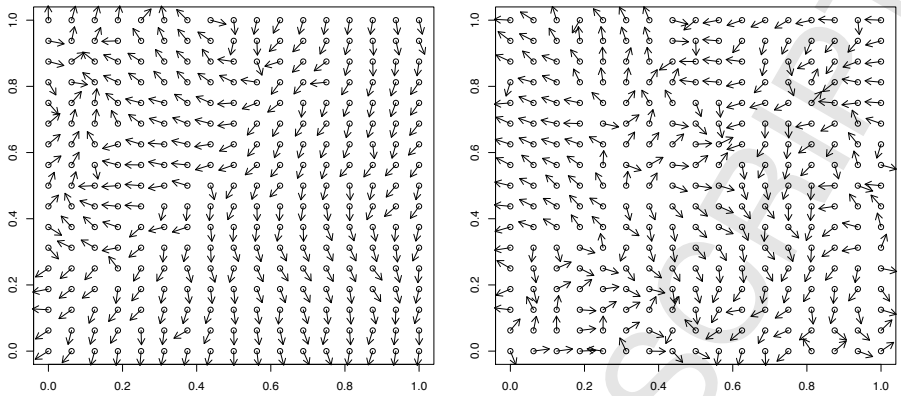


Figure 6: Simulated draws from projected Gaussian processes with different values of ϕ where (left) $\phi = 1$, signaling strong dependence and (right) $\phi = 5$, signaling weak dependence.

realizations of two projected Gaussian process using the exponential covariance function. We show strong spatial dependence in Figure 6 (left) where the spatial decay parameter is $\phi = 1$ and weak dependence in Figure 6 (right) where the spatial decay parameter is $\phi = 5$. The circles indicate the locations, the arrows indicate directions.

10. Spatial data fusion or assimilation

With spatial data, it is common to obtain information from multiple sources and at different spatial scales. For example, with environmental exposure to pollutants, we might have monitoring station data at point level, satellite data from overlapping grid cells, and computer model output at a different grid cell resolution. A similar situation arises with weather/climate data. Data fusion or assimilation refers to the objective of synthesizing multiple, often misaligned, data sources. One class of approaches can be referred to as algorithmic/pseudo-statistical. In the spirit of this article, we are interested in alternative fully model-based approaches.

Given that the data arises at different spatial scales, a second consideration is whether to scale up to a coarser scale or to scale down to a finer scale. In principle, by *averaging*, we can always scale up. Scaling down requires the assumption that, for example, a process observed at areal scale can be conceptualized at point scale. With, say, temperature measurements, we may collect average temperatures over areal units but we can envision a point level temperature surface. With, say, percent college-educated in an areal unit, there is no sensible point level percentage. Here, we assume that there is a point level process yielding a point level surface which is modeled as a realization of a Gaussian process.

We first consider scaling up, following ideas of so-called Bayesian melding (Fuentes and Raftery, 2005). We conceptualize a true exposure surface with the data sources informing about this surface. For example, if we consider particulate matter, say $PM_{2.5}$, suppose we have monitoring station data as well as computer model output data, both of which we assume will vary in some fashion around the true surface.

If $Y(\mathbf{s})$ is the true point level spatial surface, we denote the true average exposure in a grid

cell B, by $Y(B)$ which arises through the stochastic integral

$$Y(B) = \frac{1}{|B|}\int_B Y(\mathbf{s})\,d\mathbf{s}, \tag{11}$$

where $|B|$ denotes the area of grid cell B; $Y(B)$ is called the block average. Let $Z(\mathbf{s})$ denote the finitely observed surface modeled as a measurement error model. At $\mathbf{s}$, $Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s})$ where $\epsilon(\mathbf{s}) \sim N(0, \sigma_\epsilon)$. We model the true exposure process geostatistically as $Y(\mathbf{s}) = \mu(\mathbf{s}) + \eta(\mathbf{s})$ where $\mu(\mathbf{s})$ is the spatial mean surface (typically specified using meteorological covariates) and $\eta(\mathbf{s})$ is a zero mean Gaussian process.

If the second source is computer model output, we would assume that it is not calibrated. A conceptual point level output from the computer model can take the form $Q(\mathbf{s}) = \beta_0(\mathbf{s}) + \beta_1(\mathbf{s})Y(\mathbf{s}) + \delta(\mathbf{s})$ where $\beta_0(\mathbf{s})$ is additive bias, $\beta_1(\mathbf{s})$ is multiplicative bias, and $\delta(\mathbf{s})$ is white noise. We recognize this model as the very flexible spatially varying coefficients model discussed in Section 5. $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ arise from a bivariate Gaussian process using, say, coregionalization. The grid level computer model output also arises by stochastic integration, i.e., for grid cell B_j,

$$Q(B_j) = \frac{1}{|B_j|}\int_{B_j} \beta_0(\mathbf{s})\,d\mathbf{s} + \int_{B_j} \beta_1(\mathbf{s})Y(\mathbf{s})\,d\mathbf{s} + \int_{B_j} \delta(\mathbf{s})\,d\mathbf{s}. \tag{12}$$

With these first stage specifications for the point referenced monitoring data and the computer output data along with the Gaussian process second stage specification for the true exposure, we have a hierarchical model whose fitting is discussed in Fuentes and Raftery (2005). In practice, this model yields unstable model fitting. With spatially varying slopes, it emerges as too big for the data and introduces identifiability problems. So, in practice, a constant slope $\beta_1(\mathbf{s}) = \beta_1$ is adopted over the region of interest. Then, (12) simplifies $\frac{1}{|B_j|}\int_{B_j} \beta_1(\mathbf{s})Y(\mathbf{s})\,d\mathbf{s}$ to $\beta_1 Y(B_j)$. Even so, this fusion strategy becomes computationally infeasible with many grid cells, e.g., at continental scale, since every stochastic integral requires discrete approximation (Banerjee et al., 2014).

A computationally attractive alternative is downscaling. Work by Berrocal et al. (2010a,b, 2012) converts the problem to fitting a geostatistical model. Now, the monitoring station exposure at $\mathbf{s}$, $Z(\mathbf{s})$ is related to $Q(B)$, the computer model exposure for grid cell B simply as follows. For each $\mathbf{s}$ in B,

$$Y(\mathbf{s}) = \tilde{\beta}_0(\mathbf{s}) + \tilde{\beta}_1(\mathbf{s})\,x(B) + \epsilon(\mathbf{s}), \qquad \epsilon(\mathbf{s}) \overset{ind}{\sim} \mathrm{N}(0, \tau^2) \tag{13}$$

where $\tilde{\beta}_0(\mathbf{s}) = \beta_0 + \beta_0(\mathbf{s})$ and $\tilde{\beta}_1(\mathbf{s}) = \beta_1 + \beta_1(\mathbf{s})$ and $\epsilon(\mathbf{s})$ is white noise. Again, we have a spatially varying coefficients model where β_0 and β_1 provide overall calibration of the computer model and $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ provide the local adjustments. Fitting the downscaler model is computationally feasible since there are relatively few monitoring stations compared with the very large number of grid cells. Moreover, the spatially varying coefficients model is well-behaved here (Berrocal et al., 2010a).

11. Large spatial datasets

The previous sections have demonstrated a rich range of settings where Gaussian process specifications provide attractive spatial modeling. A challenge with the use of Gaussian processes arises with large datasets, which are being collected more and more frequently. In a hierarchical modeling framework or even working merely with the likelihood, evaluation requires the computation of a high dimensional inverse and determinant; with n spatial locations

inverse calculation is essentially an $O(n^3)$ operation. Since these matrices involve the unknown parameters in the covariance function, many evaluations of the likelihood will be needed. Form a computational perspective, this may prove either too costly or infeasible. Since we wish to still employ fully model-based inference, we consider Gaussian process models which can avoid this challenge. We describe two choices here, the predictive Gaussian process (Banerjee et al., 2008) and the nearest neighbor Gaussian process (Datta et al., 2016).

The predictive Gaussian process employs the idea of dimension reduction. We represent a high dimensional vector of variables in a lower dimensional space. That is, we write the $n \times 1$ vector $\mathbf{w} = (w(\mathbf{s}_1), w(\mathbf{s}_2), ..., w(\mathbf{s}_n))$ in the form $\mathbf{w} = L\tilde{\mathbf{w}}$, where $\tilde{\mathbf{w}}$ is $m \times 1$. How do we choose L and $\tilde{\mathbf{w}}$? We replace $\mathbf{w}(\mathbf{s})$ by a process $\tilde{\mathbf{w}}(\mathbf{s})$ which *projects* the original process into a finite dimensional subspace.

In fact, we use basic Gaussian kriging (Banerjee et al., 2014, Chapter 2). Consider knots $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_m^*\}$ with $m << n$. Predict $w(\mathbf{s}_0)$ based upon the variables associated with $\mathcal{S}^*$ by

$$\tilde{w}(\mathbf{s}_0) = E[w(\mathbf{s}_0)|\mathbf{w}^*] = Cov(w(\mathbf{s}_0), \mathbf{w}^*)Var(\mathbf{w}^*)^{-1}\mathbf{w}^*. \tag{14}$$

Here, $\mathbf{w}^* = \{w(\mathbf{s}_j^*), j = 1, 2, ...m\}$, $\text{Cov}(w(\mathbf{s}_0), \mathbf{w}^*) = (C(\mathbf{s}_0, \mathbf{s}_1^*; \phi), \ldots,$
$C(\mathbf{s}_0, \mathbf{s}_m^*; \phi))$, and $\text{Var}(\mathbf{w}^*) = \Sigma(\mathbf{s}_j^*, \mathbf{s}_{j'}^*; \phi)_{j,j'=1}^m$. $\tilde{w}(\mathbf{s})$ defined in this way creates a new spatial Gaussian process, referred to as the predictive process, with covariance function

$$\tilde{C}(\mathbf{s}, \mathbf{s}') = Cov(w(\mathbf{s}), \mathbf{w}^*)Var(\mathbf{w}^*)^{-1}Cov^T(w(\mathbf{s}), \mathbf{w}^*).$$

We can write $\tilde{w}(\mathbf{s}) = \ell(\mathbf{s})\mathbf{w}^*$ where $\ell(\mathbf{s}) = Cov(w(\mathbf{s}), \mathbf{w}^*)Var(\mathbf{w}^*)^{-1}$ is a $1 \times m$ vector. Then, the process realization over $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is $\tilde{\mathbf{w}} = \mathbf{L}\mathbf{w}^* \sim MVN(\mathbf{0}, \mathbf{L}\Sigma_{\mathbf{w}^*}^{-1}\mathbf{L}^T)$ with $\mathbf{L} = [\ell(\mathbf{s}_i)]_{i=1}^n$ being $n \times m$.

We fit the reduced model $Y(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\boldsymbol{\beta} + \tilde{w}(\mathbf{s}) + \epsilon(\mathbf{s}) = \mathbf{X}^T(\mathbf{s})\boldsymbol{\beta} + \ell(\mathbf{s})\mathbf{w}^* + \epsilon(\mathbf{s})$. Computations involve $\Sigma_{\mathbf{w}^*}$ $(m \times m)$ instead of $\Sigma_{\mathbf{w}}$ $(n \times n)$. This changes $O(n^3)$ computation to $O(m^3)$ computation, yielding dramatic savings when $m << n$. Model fitting with the predictive process is straightforward; we only need to introduce the $\mathbf{w}^*$.

The predictive process offers the following attractive properties: (i) it is a process itself, induced from the original process; there is no need to specify kernels or other covariance functions, (ii) it requires no restrictions to stationarity or isotropy, (iii) it immediately extends to multivariate and spatiotemporal settings, (iv) given $\mathcal{S}^*$ you *cannot* do better in a (reverse entropy) Kullback-Leibler sense with any other process of the form $\ell(\mathbf{s})\mathbf{w}^*$ (Banerjee et al., 2008), (v) when $\mathcal{S}^* = \mathcal{S}$ we recover the original model.

However, it also has several negatives: (i) with regard to knot selection, how many knots will suffice and how do we choose them? (ii) since the process lives in m-dimensional space, with more than m locations, the covariance matrix is singular; as a result, the predictive process underestimates spatial variability, putting more uncertainty into the pure error process (work of Finley et al. (2009) and Sang and Huang (2012) offers bias correction to mitigate these concerns), (iii) the process over-smoothes relative to the original Gaussian process since, unlike the original process, realizations of the predictive process are mean square differentiable of all orders.

In response to these concerns, we conclude with a very recent Gaussian process model for handling very large spatial datasets, the nearest neighbor Gaussian process (NNGP) (Datta et al., 2016). The basic idea is an extension of the work of Vecchia (1988, 1992) and Stein et al. (2004). Consider a set of reference locations $\mathcal{R} = \{\mathbf{s}_i, i = 1, 2, ..., r\}$. For a given ordering of the $\mathbf{s}_i$'s, the joint distribution of $\mathbf{w}$ can be written as $p(\mathbf{w}) = p(w(\mathbf{s}_1))p(w(\mathbf{s}_2)|w(\mathbf{s}_1))...p(w(\mathbf{s}_r)|w(\mathbf{s}_{r-1}), ...w(\mathbf{s}_1))$. The NNGP strategy is to replace the large conditioning sets with smaller, carefully chosen sets

of size at most m, where m is relatively small, e.g., a set of m nearest neighbors. So, for every $\mathbf{s}_i$, a smaller conditioning set $N(\mathbf{s}_i)$ is used to construct $\tilde{p}(\mathbf{w}) = \Pi_{i=1}^{r} p(w(\mathbf{s}_i)|\mathbf{w}_{N(\mathbf{s}_i)})$. Through suitable specification of the $N(\mathbf{s}_i)$, we can ensure that $\tilde{p}(\mathbf{w})$ is a multivariate normal density.

Now consider any set of n locations, $\mathcal{T}$. For any $\mathbf{s}$ in $\mathcal{T}$ which is in $\mathcal{R}$, we immediately have the contribution of $w(\mathbf{s})$ to the joint distribution for the $w(\mathbf{s})$ in $\mathcal{T}$. For any $\mathbf{s}$ not in $\mathcal{R}$, we contribute to the joint distribution multiplicatively, its conditional normal distribution given the m nearest neighbors in $\mathcal{R}$. The product form created in this fashion provides a valid joint multivariate normal distribution for any $\mathcal{T}$ as shown in Datta et al. (2016). Moreover, this recipe for arbitrary finite dimensional distributions creates a Gaussian process over R^2. Additionally, Datta et al. (2016) show that the associated covariance matrix for any $\mathcal{T}$ is sparse, involving only $m \times m$ matrices. Hence computation for this model becomes $O(nm^3)$. Since computation is now of linear order in n, this enables handling of very large datasets. Lastly, Datta et al. (2016) show that the performance of this process specification is not sensitive to the size of m, the ordering of the $\mathbf{s}$'s, or the selection of $\mathcal{R}$.

12. A brief summary

When implementing statistical analysis of point-referenced spatial data, a fully model-based specification enables full inference with regard to estimation and prediction, with attachment of appropriate uncertainty. The customary geostatistical model specification introduces both spatial error and pure error terms. Viewing the spatial terms as random effects suggests modeling them to be normal, bringing in a variance component. Capturing structured spatial dependence in these random effects leads to a Gaussian process specification.

However, the current geostatistical modeling world considers richer types of data and asks more demanding questions. The richer types of data include categorical and count data, multivariate data, extremes, and circular data. The more demanding questions include fusion of data sources and computation for very large datasets. We have shown here that, with suitable specification, Gaussian processes can be employed to address all of these issues. Hence, despite their somewhat innocuous appearance in the basic geostatistical model, they emerge as an invaluable tool for handling arising challenges. For point-referenced data, we are unaware of any other modeling device that provides the scope of opportunities as Gaussian processes. For us, this affirms that spatial statistics and Gaussian processes do make a beautiful marriage.

References

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2014. Hierarchical modeling and analysis for spatial data. CRC Press.

Banerjee, S., Gelfand, A.E., 2003. On smoothness properties of spatial processes. Journal of Multivariate Analysis 84, 85–100.

Banerjee, S., Gelfand, A.E., Finley, A.O., 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.

Banerjee, S., Gelfand, A.E., Sirmans, C., 2003. Directional rates of change under spatial process models. Journal of the American Statistical Association 98, 946–954.

Barry, R.P., Ver Hoef, J.M., 1996. Blackbox kriging: spatial prediction without specifying variogram models. Journal of Agricultural, Biological, and Environmental Statistics , 297–322.

Berger, J.O., De Oliveira, V., Sansó, B., 2001. Objective bayesian analysis of spatially correlated data. Journal of the American Statistical Association 96, 1361–1374.

Berrocal, V.J., Gelfand, A.E., Holland, D.M., 2010a. A bivariate space-time downscaler under space and time misalignment. The Annals of Applied Statistics 4, 1942.

Berrocal, V.J., Gelfand, A.E., Holland, D.M., 2010b. A spatio-temporal downscaler for output from numerical models. Journal of Agricultural, Biological, and Environmental Statistics 15, 176–197.

Berrocal, V.J., Gelfand, A.E., Holland, D.M., 2012. Space-time data fusion under error in computer model output: An application to modeling air quality. Biometrics 68, 837–848.

Chu, K.C., 1973. Estimation and detection for linear systems with elliptical random variables. IEEE Transactions on Automatic Control , 499–505.

Coles, S., Bawa, J., Trenner, L., Dorazio, P., 2001. An introduction to statistical modeling of extreme values. volume 208. Springer.

Cressie, N., Wikle, C.K., 2011. Statistics for spatio-temporal data. John Wiley & Sons.

Damian, D., Sampson, P.D., Guttorp, P., 2001. Bayesian estimation of semi-parametric non-stationary spatial covariance structures. Environmetrics 12, 161–178.

Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E., 2016. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. Journal of the American Statistical Association, Series B (Statistical Methodology) .

Diggle, P.J., Tawn, J., Moyeed, R., 1998. Model-based geostatistics. Journal of the Royal Statistical Society: Series C (Applied Statistics) 47, 299–350.

Duan, J.A., Guindani, M., Gelfand, A.E., 2007. Generalized spatial Dirichlet process models. Biometrika 94, 809–825.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression: models, methods and applications. Springer Science & Business Media.

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

Fuentes, M., Raftery, A.E., 2005. Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. Biometrics 61, 36–45.

Gelfand, A.E., 2012. Hierarchical modeling for spatial data problems. Spatial Statistics 1, 30–39.

Gelfand, A.E., Ghosh, S., 2013. Hierarchical modeling, in: Damien, P., Dellaportas, P., Polson, N., Stephens, D. (Eds.), Bayesian Theory and Applications: a Tribute to Adrian Smith. Oxford University Press. chapter 3, pp. 33–49.

Gelfand, A.E., Kim, H.J., Sirmans, C., Banerjee, S., 2003. Spatial modeling with spatially varying coefficient processes. Journal of the American Statistical Association 98, 387–396.

Gelfand, A.E., Kottas, A., MacEachern, S.N., 2005. Bayesian nonparametric spatial modeling with Dirichlet process mixing. Journal of the American Statistical Association 100, 1021–1035.

Gelfand, A.E., Schmidt, A.M., Banerjee, S., Sirmans, C., 2004. Nonstationary multivariate process modeling through spatially varying coregionalization. Test 13, 263–312.

Higdon, D.M., Anderson, C., Barnett, V., Chatwin, P.C., El-Shaarawi, A.H., 2002. Space and space-time modeling using process convolutions. Quantitative Methods for Current Environmental Issues 3754, 37–56.

Jammalamadaka, S.R., Sengupta, A., 2001. Topics in circular statistics. volume 5. World Scientific.

Jaworski, P., Durante, F., Hardle, W.K., Rychlik, T., 2010. Copula theory and its applications. Springer.

Jona Lasinio, G., Gelfand, A., Jona Lasinio, M., 2012. Spatial analysis of wave direction data using wrapped gaussian processes. The Annals of Applied Statistics 6, 1478–1498.

Kim, H.M., Mallick, B.K., 2003. A note on Bayesian spatial prediction using the elliptical distribution. Statistics & Probability Letters 64, 271–276.

Kottas, A., Krnjajić, M., 2009. Bayesian semiparametric modelling in quantile regression. Scandinavian Journal of Statistics 36, 297–319.

Kotz, S., Kozubowski, T.J., Podgórski, K., 2001. Asymmetric multivariate laplace distribution, in: The Laplace Distribution and Generalizations. Springer, pp. 239–272.

Lum, K., Gelfand, A.E., 2012. Spatial quantile multiple regression using the asymmetric laplace process. Bayesian Analysis 7, 235–258.

Mastrantonio, G., Gelfand, A.E., Lasinio, G.J., 2015. The wrapped skew Gaussian process for analyzing spatio-temporal data. Stochastic Environmental Research and Risk Assessment , 1–12.

Meiring, W., Monestiez, P., Sampson, P., Guttorp, P., 1997. Developments in the modelling of nonstationary spatial covariance structure from space-time monitoring data. Geostatistics Wollongong 96, 162–173.

Minozzo, M., Ferracuti, L., 2012. On the existence of some skew-normal stationary processes. Chilean Journal of Statistics 3, 157–170.

Moller, J., Waagepetersen, R.P., 2003. Statistical inference and simulation for spatial point processes. CRC Press.

Nelsen, R.B., 2013. An introduction to copulas. volume 139. Springer Science & Business Media.

Padoan, S.A., Ribatet, M., Sisson, S.A., 2010. Likelihood-based inference for max-stable processes. Journal of the American Statistical Association 105, 263–277.

Palacios, M.B., Steel, M.F.J., 2006. Non-gaussian bayesian geostatistical modeling. Journal of the American Statistical Association 101, 604–618.

Petrone, S., Guindani, M., Gelfand, A.E., 2009. Hybrid Dirichlet mixture models for functional data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71, 755–782.

Ramsay, J.O., Silverman, B.W., 2002. Applied functional data analysis: methods and case studies. volume 77. Springer New York.

Rodriguez, A., Kottas, A., 2014. Bayesian Nonparametric Mixture Models: Methods and Applications. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.

Rue, H., Held, L., 2005. Gaussian Markov random fields: theory and applications. CRC Press.

Sang, H., Gelfand, A.E., 2009. Hierarchical modeling for extreme values observed over space and time. Environmental and Ecological Statistics 16, 407–426.

Sang, H., Gelfand, A.E., 2010. Continuous spatial process models for spatial extreme values. Journal of Agricultural, Biological, and Environmental Statistics 15, 49–65.

Sang, H., Huang, J.Z., 2012. A full scale approximation of covariance functions for large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74, 111–132.

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

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

Vecchia, A.V., 1992. A new method of prediction for spatial regression models with correlated errors. Journal of the Royal Statistical Society. Series B (Methodological) , 813–830.

Villani, M., Larsson, R., 2006. The multivariate split normal distribution and asymmetric principal components analysis. Communications in Statistics - Theory and Methods 35, 1123–1140.

Wackernagel, H., 2003. Multivariate Geostatistics. Springer Science & Business Media.

Wang, F., Gelfand, A.E., 2013. Directional data analysis under the general projected normal distribution. Statistical methodology 10, 113–127.

Wang, F., Gelfand, A.E., 2014. Modeling space and space-time directional data using projected Gaussian processes. Journal of the American Statistical Association 109, 1565–1580.

West, B.T., Welch, K.B., Galecki, A.T., 2014. Linear mixed models: a practical guide using statistical software. CRC Press.

Yaglom, A., 1987. Correlation Theory of Stationary and Related Random Functions. Springer-Verlag.

Zhang, H., El-Shaarawi, A., 2010. On spatial skew-Gaussian processes and applications. Environmetrics 21, 33–47.