On generating a flexible class of non-stationary spatial models using Gaussian predictive processes

Sujit K. Sahu*and Sabyasachi Mukhopadhyay,
University of Southampton, UK

January 13, 2016

Abstract

This article proposes a flexible class of non-stationary spatial models by using recently developed Gaussian predictive processes. So far these processes are only used as approximate dimension reduction models for analysing large spatial data sets. The contribution of the current article lies in proposing these models even for small sample sizes and studying the nature of non-stationarity implied by these predictive processes under various scenarios of selection of the knot locations where the predictive process is to be anchored for both small and large data sets. Results obtained here show that different random and non-random choices of knot-locations lead to new flexible forms of non-stationary covariance functions not yet studied in the literature. These new covariance functions give rise to new flexible Bayesian predictive models but do not complicate the fitting and analysis methods unlike other non-stationary models. The proposed methods are illustrated using two practical data sets on modelling air pollution exposure in London and the other on modelling a well-known data set on scallop abundance in the Atlantic Ocean near the City of New York.

*E-mail: S.K.Sahu@soton.ac.uk.

1 Introduction

Stochastic spatial models based on Gaussian processes are experiencing a surge of popularity in recent literature due to their abilities to investigate spatial variation in many physical quantities of interest in diverse application areas. A stationary Gaussian process with an isotropic covariance function is often the default choice for statistical modellers since such an assumption implies a tractable model leading to easily amenable analysis and computation. The Gaussian processes, used naively, lead to stationary and isotropic covariance models for data. Due to their analytical tractability, these models are not only convenient to specify but also are easy to fit and analyse using contributed software packages inside the `R` language environment. Recent references include: Banerjee *et al.* (2015); Cameletti *et al.* (2013); Cressie and Wikle (2011); Finley *et al.* (2015) and Bakar and Sahu (2015).

Simplicity of the Gaussian processes, however, does not often represent reality and practical data sets often exhibit non-stationarity. Non-stationarity arises when the distribution, or its features, e.g., means, variances, and correlations, depend on the actual spatial locations where random variables have been observed. Non-stationarity is hard to generalise into a simplistic parametric model although there is a relatively large literature on constructing non-stationary models using deformation, see e.g. (Sampson and Guttorp, 1992; Schmidt and O'Hagan, 2003) and kernel mixing, see e.g. (Higdon, 1998; Paciorek and Schervish, 2006); Section 3.2 of Banerjee *et al.* (2015) provides a review. More recent articles in this area include: Konomi *et al.* (2014) who use a non-stationary covariance function constructed based on adaptively selected partitions; Guhaniyogi *et al.* (2013) who use spatially varying cross-covariance models; Katzfuss (2013) who uses spatial basis functions with nonstationary Matérn covariance functions. Alternatives to these approaches are those based on approximations using stochastic partial differential equations, see e.g. (Lindgren *et al.*, 2011; Bolin and Lindgren, 2011; Ingebrigtsen *et al.*, 2014) and Fuglstad *et al.* (2015) where non-stationary models have been proposed by locally varying co-efficients in the approximations.

The main objective of this paper is to introduce a method to generate flexible non-stationary spatial models which are also based on Gaussian processes – so that it is also easy to fit and predict with the models. The generating mechanism relies on specification of a Gaussian predictive process (GPP), see Banerjee *et al.* (2008), which induces non-stationarity. However, there are several un-resolved questions regarding the use of GPP as a method to generate non-stationary spatial models. What are the covariance properties of the new spatial process induced by GPP? How do those properties change as the *knots* are moved around in the study region? How does a

particular clustering of the knots affect these properties? What happens if the knots are instead specified randomly according to a specific point pattern model, see e.g. Guhaniyogi *et al.* (2011)? How does the choice of the number of knots influence the covariance structure of the new process?

Our main contribution here is to investigate the above issues in detail with practical examples in order to develop accurate predictive models. Here we find that the GPP defines a new class of flexible spatial models which are able to capture non-stationarity as yet un-explored in the literature. The new processes generate models having non-stationarity in both the marginal variances and correlations. The flexible nature of non-stationarity is controlled by the number as well as the positioning of the knots. The flexibility ranges from complete stationarity, corresponding to having no knots at all, to highly non-stationary models corresponding to a dense distribution of the knots covering the entire study region.

As expected, the generated non-stationary models produce anisotropic covariance functions (Zimmerman, 1993; Ecker and Gelfand, 1999). Anisotropic covariance functions are those for which the covariance depends not only on the distance but also on the direction between random observations at any two locations. For example, the covariance function as a function of distance may decay at different rates when measured at different directions (range anisotropy). It is also be possible to have sill anisotropy where the covariance function may asymptote to different levels when measured at different directions. Zonal anisotropy is generated by placing knots with different clustering patterns at different zones of the study region. In addition, the generated processes will exhibit geometric anisotropy if the parent Gaussian process also posses the same characteristic. In this sense, the generated processes are capable of generating anisotropic covariance functions which can be sill, range, geometric and zonal at the same time.

In practical and empirical data modelling situations, it is often difficult to decide a-priori what type of anisotropy (or non-stationarity) will be the most appropriate unless there is specific information regarding the processes to be modelled. Bayesian modelling experimentation with different random and non-random schemes for knot placements is proposed to be a possible solution to this problem. Competing Bayesian models are to be compared using either Bayesian predictive model choice criteria, e.g., Gelfand and Ghosh (1998); Spiegelhalter *et al.* (2002), or empirical validation measures calculated using predictions for set aside observations. The paper details setting up of the full Bayesian model corresponding to random placement of the knots within the study region and then compares these models with those specified using default non-random designs for knot placements.

The plan of the remainder of the paper is as follows. In Section 2 we review the predictive process model and explore different ways to generate anisotropic models. Section 2.1 illustrates the nature of anisotropy generated by the proposed method by two theoretical examples. Full Bayesian hierarchical models and prediction details based on the anisotropic covariance functions are laid out in Section 3. Section 4 contains illustrations of the methods using two practical examples: one for the scallop catch data set and the other for modelling NO_2 pollution levels in the city of London in 2011. A few summary remarks are placed in Section 5.

2 GPP method for generating non-stationary models

A GPP is simply defined as the process induced by Kriging. To formally define this, assume that $w(\mathbf{s})$ is the spatial random effect at a location $\mathbf{s}$ and it follows a zero-mean stationary GP with an isotropic covariance function $\sigma_w^2 C(\cdot)$ where $C(\cdot)$ is assumed to be a member of the Matérn family. The correlation function, $C(\cdot)$, will depend on two additional parameters: smoothness ν and the rate of decay ϕ but these are suppressed from the notation C for convenience.

Given a set of m point locations $\mathbf{S}_m^* = (\mathbf{s}_1^*, \ldots, \mathbf{s}_m^*)$, which are to be called the *knot-locations* or simply the *knots*, in a d-dimensional study region D (D is a subspace of $\mathbb{R}^d$), the GPP at a new location $\mathbf{s}$, denoted by $\tilde{w}(\mathbf{s})$, is defined as the conditional expectation of the GP $w(\mathbf{s})$ given the m realisations at the knots denoted by $\mathbf{w}^* = (w(\mathbf{s}_1^*), \ldots, w(\mathbf{s}_m^*))^T$. In particular,

$$\tilde{w}(\mathbf{s}) = E\left[w(\mathbf{s})|\mathbf{w}^*\right]. \tag{1}$$

Properties of the underlying GP yield that: (i) marginally $\mathbf{w}^*$ follows $N\left(\mathbf{0}, \sigma_w^2 S_{w^*}\right)$ where S_{w^*} is the $m \times m$ correlation matrix whose entries are formed using $C(\cdot)$ and (ii) the random vector $(w(\mathbf{s}), \mathbf{w}^*)^T$ follows the multivariate Gaussian distribution with mean zero and a covariance matrix given by $\sigma_w^2 \begin{pmatrix} 1 & \mathbf{c}^{*T}(\mathbf{s}) \\ \mathbf{c}^*(\mathbf{s}) & S_{w^*} \end{pmatrix}$ where $\mathbf{c}^*(\mathbf{s}) = (C(|\mathbf{s} - \mathbf{s}_1^*|), \ldots, C(|\mathbf{s} - \mathbf{s}_m^*|))^T$. Multivariate Gaussian theory yields that

$$\tilde{w}(\mathbf{s}) = \mathbf{c}^{*T}(\mathbf{s}) S_{w^*}^{-1} \mathbf{w}^*. \tag{2}$$

Consider the following simplification of the above setting that motivates the central issue in the paper. Suppose that the real line $\mathbb{R}^1$ is the study region D. With one knot point ($m = 1$), s_1^*, say at the origin, and assuming the exponential correlation

function with decay parameter $\phi > 0$ in the GP with unit spatial variance we can easily see that $\tilde{w}(s) = \exp(-\phi\,|s|)w^*(0)$ according to (2) where $w^*(0) \sim N(0, \sigma_w^2)$. The marginal variance given by $\sigma_w^2 \exp(-2\phi|s|)$ depends on the location through $|s|$ hence this will generate a non-stationary process. In general, the covariance between $\tilde{w}(s)$ and $\tilde{w}(s')$ will depend not only on the distance $|s - s'|$ but also on the relative positioning of s and s' with respect to the origin, the sole knot-location here. However, in this one dimensional example the correlation between $\tilde{w}(s)$ and $\tilde{w}(s')$ will be equal to one due to the dimension reduction performed by the GPP for any $s \neq s'$. To avoid this degeneracy caused by dimension reduction, m must be taken to be greater than n, or another independent process must be added for data modelling as discussed in Section 3.

Further complexities in the covariance function are easily introduced by: (i) assuming specific clustering processes for the knot-locations and (ii) assuming the knot locations to be assigned at random over a finite subspace in D as we demonstrate below for a two-dimensional example.

Returning to the general GPP (2), for two locations $\mathbf{s}$ and $\mathbf{s}'$ the covariance between $\tilde{w}(\mathbf{s})$ and $\tilde{w}(\mathbf{s}')$ is given by:

$$\text{Cov}\,(\tilde{w}(\mathbf{s}), \tilde{w}(\mathbf{s}')) \equiv \sigma_w^2\,\tilde{C}(\mathbf{s}, \mathbf{s}') = \sigma_w^2\,\mathbf{c}^{*T}(\mathbf{s}) S_{w^*}^{-1} \mathbf{c}^*(\mathbf{s}'). \tag{3}$$

The above defines a valid non-negative definite covariance function since according to (2), $\tilde{w}(\mathbf{s})$ and $\tilde{w}(\mathbf{s}')$ are finite linear combinations of the elements of $\mathbf{w}^*$, which is equipped with a valid positive definite covariance function $C(\cdot)$. As discussed in the above example, $\tilde{C}(\cdot, \cdot)$ will give rise to a singular correlation matrix when the number of realisations n is larger than m due to the dimension reduction. The singularity will not arise when we explore the correlation structure in the remainder of this section with $n = 2$ points, $\mathbf{s}$ and $\mathbf{s}'$, and by taking $m \geq 2$. Of-course, in practical data modelling settings, non-singular models are guaranteed by adopting the $\tilde{w}(\cdot)$ process in a hierarchical model as in Section 3.

Clearly, $\tilde{C}(\mathbf{s}, \mathbf{s}')$ depends on both $\mathbf{s}$ and $\mathbf{s}'$ and not only through the separation vector $\mathbf{s} - \mathbf{s}'$ or the distance $|\mathbf{s} - \mathbf{s}'|$. As a result, the model specification with $\tilde{w}(\mathbf{s})$ as the spatial effects will also imply non-stationary and, hence, anisotropic correlation structure.

The covariance function (3) is easily used to define the traditional semivariogram,

denoted by, $\tilde{\gamma}(\mathbf{s}, \mathbf{h})$ of the spatial effects $\tilde{w}(\mathbf{s})$ as follows:

$$\begin{aligned} 2\tilde{\gamma}(\mathbf{s}, \mathbf{s}') &= \mathrm{Var}\left[\tilde{w}(\mathbf{s}) - \tilde{w}(\mathbf{s}')\right] \\ &= E\left[\tilde{w}(\mathbf{s}) - \tilde{w}(\mathbf{s}')\right]^2 \\ &= E\left[\mathbf{c}^{*T}(\mathbf{s}) S_{w^*}^{-1} \mathbf{w}^* - \mathbf{c}^{*T}(\mathbf{s}') S_{w^*}^{-1} \mathbf{w}^*\right]^2 \\ &= E\left[\left(\mathbf{c}^{*T}(\mathbf{s}) - \mathbf{c}^{*T}(\mathbf{s}')\right) S_{w^*}^{-1} \mathbf{w}^*\right]^2 \\ &= \left(\mathbf{c}^{*T}(\mathbf{s}) - \mathbf{c}^{*T}(\mathbf{s}')\right) S_{w^*}^{-1} \left(\mathbf{c}^*(\mathbf{s}) - \mathbf{c}^*(\mathbf{s}')\right). \end{aligned}$$

Non-stationarity of the $\tilde{w}(\mathbf{s})$ is also apparent from the dependence of the semivariogram, $\tilde{\gamma}(\mathbf{s}, \mathbf{s}')$ on both $\mathbf{s}$ and $\mathbf{s}'$. Further exploration of non-stationarity is proceeded with the covariance function (3) itself instead of the semivariogram, $\tilde{\gamma}$ for ease of interpretation since $\tilde{C}$ uniquely determines $\tilde{\gamma}$ but not the converse. Moreover, for Gaussian processes covariance functions are natural quantities to look at and in practical modelling situations those are the ones that must be specified but not derived quantities like the semivariogram.

Exploration of non-stationarity is not straightforward because of the dependence of $\tilde{C}(\mathbf{s}, \mathbf{s}')$ on both of its arguments and not only on the separation vector $\mathbf{h} = \mathbf{s} - \mathbf{s}'$ or the distance $|\mathbf{h}|$. However, to facilitate comparison with (or departure from) stationarity we first write it as a function of $\mathbf{s}$, $|\mathbf{h}|$ and also of the angle at which $\mathbf{s}'$ lies with respect to the reference axes used to define the underlying GP $w^*(\mathbf{s})$. The covariance function still will vary with $\mathbf{s}$, which itself can be any point within the study region, D.

To study the nature of non-stationarity in the $\tilde{w}$ process, we consider variation of $\tilde{C}(\mathbf{s}, \mathbf{s}')$ by taking $\mathbf{s}$ as a 'central' location within D. In practical modelling situations when data locations are available, we can take the unique centroid of the convex hull of all the locations. This allows us to study the induced directional correlation structure as data locations move away from the centroid, denoted by $\mathbf{s}^{**}$, to the boundaries of D. The central location, to be used for exploration purposes only and not for inference, can also be an 'externality' in the study region, for example, the centre of the business district when modelling land prices, see Banerjee *et al.* (2015).

The number, m, and configuration of the knots play a major role in dictating the nature of non-stationarity as the examples below will illustrate. A novel proposal here is to allow the knot-locations to be random given m. This randomness generates further flexibility in modelling and is the preferred approach as developed and illustrated in the later sections. Before going into the specific modelling and computing details, we first note that the correlation function $\tilde{C}(\mathbf{s}, \mathbf{s}')$ will be a random quantity if the knot locations $\mathbf{S}_m^*$ are also random. Hence, by using the covariance identity

$$\mathrm{Cov}(X, Y) = E\mathrm{Cov}(X, Y|Z) + \mathrm{Cov}(E(X|Z), E(Y|Z))$$

for any three random variables, X, Y and Z, we obtain

$$\tilde{C}(\mathbf{s},\mathbf{s}') = E_{\pi(m),\pi(\mathbf{S}^*_m)} E\left[\mathbf{c}^{*T}(\mathbf{s}) S^{-1}_{w^*} \mathbf{c}^*(\mathbf{s}') | m, \mathbf{S}^*_m\right] \tag{4}$$

where we continue to use $\tilde{C}(\mathbf{s},\mathbf{s}')$ to denote the expected covariance and the outer expectation is taken over the distributions of m and $\mathbf{S}^*_m$ denoted respectively by $\pi(m)$ and $\pi(\mathbf{S}^*_m)$. Note that the second term in the covariance identity vanishes because the conditional expectation of $\tilde{w}(\mathbf{s})$ given m and $\mathbf{S}^*_m$ is zero since $E(\mathbf{w}^*) = \mathbf{0}$ since it is a random realisation of the underlying zero mean GP.

We use Monte Carlo sampling to evaluate the outer expectation in (4) as follows. At the ℓth Monte Carlo replication (out of L where L is large) we generate an m_ℓ from $\pi(m)$ and and a set of m_ℓ random knots $\mathbf{S}^*_{m_l}$ from $\pi(\mathbf{S}^*_{m_\ell})$ and using those values evaluate $\tilde{C}^{(\ell)}(\mathbf{s},\mathbf{s}') = \mathbf{c}^{*(\ell)}(\mathbf{s})^T S^{-1(\ell)}_{w^*} \mathbf{c}^{*(\ell)}(\mathbf{s}')$ where $\mathbf{c}^{*(\ell)}(\mathbf{s})$ is obtained by plugging in the values of m_ℓ and $\mathbf{S}^*_{m_l}$ in $\mathbf{c}^*(\mathbf{s})$. The matrix $S^{(\ell)}_{w^*}$ is also obtained similarly. The $\tilde{C}^{(\ell)}(\mathbf{s},\mathbf{s}'), \ell = 1,\ldots,L$ values are averaged to obtain an estimate of the true covariance $\tilde{C}(\mathbf{s},\mathbf{s}')$. In our implementation, we have taken L to be 500, but taking a larger value did not change the reported conclusions.

Thus the most general method we propose to generate a nonstationary process is based on a random number of knots which are also selected according to a random point pattern distribution over the study region of interest D. In addition, we also propose a number of intermediate strategies ranging from this random allocation of m knots to a fixed space-filling design to generate flexible models. These methods are illustrated in the next section with two examples one each on one and two dimensional sub-spaces.

2.1 Illustration of non-stationarity and anisotropy

Assume that the study region D is the compact region, $[-a,a] \times [-a,a]$ where we take $a = 2$ for illustration purposes. Origin is taken as the centroid and we work with orthogonal axes. We examine the correlation between two realisations, one at the origin and the other at a unit distance away from the origin but at different angles starting at 0 and by taking increments of 45^0, i.e., 0, 45, 90, 135, 180, 225, 270 and 315. We take $m = 100$ knot locations within the domain. We adopt the exponential correlation function with a fixed decay parameter of 0.2 for the underlying GP. To study the induced correlation structure we consider several designs for the knot points.

Our first two designs are taken to be regular grids one of which spans the entire study region and the other within a smaller rectangle containing the origin, see the two plots in the first column of Figure 1. The two plots in the second column show the covariance as a function of distance away from the origin. The covariance

curves collapse to just two in both cases because of the complimentary angles used to calculate those and the periodicity of the trigonometric functions for complimentary angles. The plots show sill and range anisotropy since the covariance function is seen to decay to zero at a different rate of the distance when that is calculated at different angles, see e.g. Zimmerman (1993) and Section 2.2 of Banerjee *et al.* (2015)) for formal definitions.

More interesting type of anisotropy arises when the knot-design is taken as random, see Figure 2. In the first design the knots points have been placed completely at random and in the second case the knots are placed at random but all within the first quadrant to see the effect of clustering of the knots in one particular sub-region which may have been incorrectly chosen by the modeller. The covariance function plots show a full spectrum of range and sill anisotropy and show the flexibility of the GPP method to generate anisotropic and hence non-stationary models. Here edge effects can be clearly seen in Figure 2, whereby the correlations are not asymptotically reduced to zero. Clearly, correlations must be calculated for larger distances to see the asymptotics. A different random realisation of the knots will introduce a different anisotropic model.

Figure 3 illustrates the effect of varying m, the knot size on anisotropy and non-stationarity. Here the correlation function is plotted for m (=10, 50 and 100) randomly selected knots when distances are calculated along the perpendicular line. The other parameters and settings are chosen exactly as in panel (b) of Figure 2. The plot of the correlation function in the isotropic case without any knots is also provided for comparison purposes. The plot shows a variety of possible non-stationary correlation functions without any possible ordering of the correlation curves which arises due to the random placing of different number of knot points.

Though the GPP method offers a flexible class of anisotropic models, it is difficult to get a mathematical form of the covariance function from the general form in (3). However, from the equation (3) it is clear that the main parameters determining the shape of the covariance function are the number of knots m, the form of the correlation function and the relative arrangement of the knots $\mathbf{S}^*$ with respect to the points of interest $\boldsymbol{s}$. The complexity of the covariance form in(3) arises mainly because it is not straightforward to determine how a given choice of knots and other parameters affect the shape of the correlation curve. To illustrate the covariance function we consider the following example.

Suppose that the study region is the two dimensional plain $\mathbb{R}^2$, and there are only two knots. One of the knots lies in the first quadrant and the other is placed in the third quadrant of the orthogonal coordinate axes. We denote the knots as (h, h) and

$(-h, -h)$, where $h > 0$ is the pre-fixed value (shown as red points for several values of h in Figure 4). The points of interest $\boldsymbol{s}$ are chosen as $\boldsymbol{s}_i^E = \frac{i}{\sqrt{2}} \times 0.08 \times h(1,1)$ along the North-East (NE) direction and $\boldsymbol{s}_i^W = \frac{i}{\sqrt{2}} \times 0.08 \times h(-1,1)$ along the North-West (NW) direction, $i = 1, \ldots, 1000$, see Figure 4 for example locations.

A closed form expression for the correlation function in (3) is obtained when the origin, denoted by $\mathbf{s}_0$, is one of the two locations of interest. In fact, we have

$$\tilde{C}(\mathbf{s}_i, \mathbf{s}_0) = \exp(-2\phi d_{i1} h) + \exp(-\phi h(d_{i1} + d_{i2} + 2\sqrt{2})) + \exp(-2\phi d_{i2} h), \qquad (5)$$

where $d_{ik} = d(\mathbf{s}_i, \mathbf{s}_k^*)/h$ and $d(\mathbf{s}_i, \mathbf{s}_k^*)$ is the distance between $\mathbf{s}_i$, $i = 1, \ldots, n$ and the knot $\mathbf{s}_k^*$, for $k = 1, 2$. Here $\mathbf{s}_i$ can be either $\boldsymbol{s}_i^E$ or $\boldsymbol{s}_i^W$. We illustrate this covariance function by constructing a two dimensional surface plot by varying h and i for $\mathbf{s}_i = \mathbf{s}_i^E$ in the NE-direction as follows. Figure 5 shows the surface plot of $\tilde{C}(\mathbf{s}_i, \mathbf{s}_0)$ where $\mathbf{s}_i = \boldsymbol{s}_i^E = \frac{i}{\sqrt{2}} \times 0.08 \times h(1,1)$ for $i = 1, \ldots, 1000$ and then taking 100 equispaced values of h between 0.01 and 1. As h and i increase the non-linearity of the correlation functions decrease. The anisotropic behaviour of the correlation function tends to change towards isotropy as h and i increase which also implies that the relative distance between the knots and s_i^E increases.

The above form (5) indicates presence of different forms of anisotropy, e.g. range, angular and nugget, in the model, see Zimmerman (1993) for definitions and numerical illustrations. The GPP based models developed here are much easier to fit in practical situations as we shall discuss in the next section.

3 Hierarchical model specification

Our starting point of spatial modelling is an assumed data realisation of the random variable $\boldsymbol{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$ at n-locations, $\mathbf{s}_1, \ldots, \mathbf{s}_n$, which we assume not to be preferentially sampled, see e.g. Gelfand *et al.* (2012). Also assume that there are p-covariates, $\boldsymbol{x}(\mathbf{s})$ measured along with $Z(\mathbf{s})$ at each data and prediction site $\mathbf{s}$. A spatial random-effect model with nugget effect, see e.g. Cressie and Wikle (2011) is given by:

$$Z(\mathbf{s}) = \boldsymbol{x}(\mathbf{s})^T \boldsymbol{\beta} + \tilde{w}(\mathbf{s}) + \epsilon(\mathbf{s}) \qquad (6)$$

where $\boldsymbol{\beta}$ denotes the unknown regression coefficients and $\epsilon(\mathbf{s}) \sim N(0, \sigma_\epsilon^2)$ is the nugget effect measuring micro-scale variation around $\mathbf{s}$ and is independent across locations and also independent of $\tilde{w}(\mathbf{s})$.

The full Bayesian hierarchical model is specified is as follows. As in Guhaniyogi *et al.* (2011), we allow the m-knots $\mathbf{S}_m^*$ to be random according to a non-homogeneous

Poisson point process model with an assumed intensity function $\lambda(\mathbf{s})$ so that

$$\pi\left(\mathbf{S}_m^*\right) = \left(\lambda(D)\right)^{-m} \prod_{j=1}^{m} \lambda(\mathbf{s}_j),$$

where $\lambda(D) = \int_D \lambda(\mathbf{s}) d\mathbf{s}$. There are many possibilities for choosing the intensity function $\lambda(\mathbf{s})$. For example, one can assume spatially varying explanatory variables, $\boldsymbol{q}(\mathbf{s})$ say, to inform the intensity, i.e. $\log(\lambda(\mathbf{s})) = \boldsymbol{q}(\mathbf{s})^T\boldsymbol{\gamma}$ where $\boldsymbol{\gamma}$ are unknown parameters. In a similar vein, Guhaniyogi *et al.* (2011) propose that

$$\log(\lambda(s)) = \frac{1}{m}\sum_{j=1}^{m} N_2\left(\mathbf{s} \mid \mathbf{u}_j, \Sigma_\lambda\right),$$

where $N_2\left(\mathbf{s} \mid \mathbf{u}_j, \Sigma_\lambda\right)$ denotes the density, evaluated at $\mathbf{s}$, of the bivariate normal distribution with unknown mean $\mathbf{u}_j$ and covariance matrix Σ_λ. These unknown parameters are then proposed to be estimated using the full Bayesian model which is completed by assuming suitable prior distributions for them. However, we can avoid this extra level of parametric uncertainty by discretizing the study region as follows.

We envision that there are M total number of possible knot locations denoted by $\mathbf{s}_1^*, \ldots, \mathbf{s}_M^*$ each having an associated probability of selection

$$\pi(\mathbf{s}_j^*) = \frac{p(\mathbf{s}_j^*)}{\sum_{j=1}^{M} p(\mathbf{s}_j^*)} \tag{7}$$

where $p(\mathbf{s}_j)$ is thought to provide a covariate like information for selecting the knots. For example, we may use a population density surface in an environmental monitoring situation that will guarantee knots being placed at high density areas. We propose sampling without replacement to avoid duplicated knots.

Conditional on m we assume the GPP specification (2) given by $\tilde{w}(\mathbf{s}) = \mathbf{c}^{*T}(\mathbf{s})S_{w^*}^{-1}\mathbf{w}^*$ where $\mathbf{w}^*$ is a realisation of the underlying zero mean Gaussian process with spatial variance σ_w^2 and isotropic Matérn correlation function $C(\cdot;\nu,\phi)$ where ν and ϕ are the smoothness and the decay parameter respectively.

The Bayesian model is completed by assuming suitable prior distributions for all the parameters and the hyper-parameters. As is often used, we shall assume normal prior distribution with zero mean and large variance, 10^4 say, for the regression parameter $\boldsymbol{\beta}$. For the variance components σ_ϵ^2 and σ_w^2 we assume that their inverses follow the Gamma distribution with parameters a and b, which we take to be 2 and 1 respectively. These values imply a proper prior distribution for each of the two variance components and experimentation here shows that inference is not sensitive to these choices.

The logarithm of the full posterior distribution is given by:

$$\begin{aligned}\log\left(\pi\left(m, \mathbf{S}_m^*, \mathbf{w}\left(\mathbf{S}_m^*\right), \boldsymbol{\theta}|\mathbf{z}\right)\right) &\propto -\tfrac{n}{2}\log(\sigma_\epsilon^2) \\ &- \tfrac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\left(z(\mathbf{s}_i) - \boldsymbol{x}(\mathbf{s}_i)^T\boldsymbol{\beta} - \tilde{w}(\mathbf{s}_i)\right)^2 \\ &- m\log(\lambda(D)) + \textstyle\sum_{j=1}^{m}\log(\lambda(\mathbf{s}_j)) \\ &- \tfrac{m}{2}\log(\sigma_w^2) - \tfrac{1}{2}\log|S_w| - \tfrac{1}{2\sigma_w^2}(\mathbf{w}^*)^T S_w^{-1}\mathbf{w} \\ &+ \log(\pi(\boldsymbol{\theta}))\end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_w^2, \nu, \phi)^T$ and $\pi(\boldsymbol{\theta})$ denotes the prior distribution of $\boldsymbol{\theta}$. Implementing the Gibbs sampler with Metropolis-Hastings steps is straightforward, see e.g. Section 3.2 of Guhaniyogi *et al.* (2011). Our implementation differs from theirs only when updating the knot-locations $\mathbf{S}_m^*$. Discretization of the space with M possible grid locations allows us to avoid having to evaluate the integral for $\lambda(D)$. Consequently, to update $\mathbf{S}_m^*$ we can simply simulate m proposed knots from the prior (7) without replacement and then use a Metropolis-Hastings step to accept the proposed knots. Alternatively, conditional on m, to update $\mathbf{S}_m^*$, we can find a new set of m proposal knots based on the current set by shifting each knot according to a random walk centred around the corresponding current knot. The proposed set of knots is then accepted using the appropriate Metropolis-Hastings step. Acceptance rate of this scheme is dependent on the step size of the random walk and is tuned to have about 30% (Gelman *et al.*, 1996). The starting configuration of the knots is taken to be according to a space filling design.

Predicting the response $Z(\mathbf{s}_0)$ at a new location $\mathbf{s}_0$ is achieved by the posterior predictive distribution

$$\pi(z(\mathbf{s}_0)|\mathbf{z}) = \int \pi(z(\mathbf{s}_0)|m, \mathbf{S}_m^*, \mathbf{w}^*, \boldsymbol{\theta}, \mathbf{z})\pi(m, \mathbf{S}_m^*, \mathbf{w}^*, \boldsymbol{\theta}|\mathbf{z})dmd\mathbf{S}_m^*d\mathbf{w}^*d\boldsymbol{\theta}$$

where $\boldsymbol{\theta}$ denotes the parameter vector $(\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_w^2, \nu, \phi)^T$. Note that since m is assumed to be discrete, integrating m out in the above must be taken as an appropriate summation. MCMC samples from the posterior distribution facilitate evaluation of the above predictive distribution. Here $z(\mathbf{s}_0)$ is assumed to be independent of $\mathbf{z}$ according to the top level model given all the parameters and the realisation of the GP. Now $\pi(z(\mathbf{s}_0)|m, \mathbf{S}_m^*, \mathbf{w}^*, \boldsymbol{\theta})$ requires $\tilde{w}(\mathbf{s}_0)$ which is calculated as $\mathbf{c}^*(\mathbf{s}_0)S_{w^*}^{-1}\mathbf{w}^*$, continuing to use (1). At the jth MCMC iteration with a posterior sample of $m^{(j)}, \mathbf{S}_m^{*(j)}, \mathbf{w}(\mathbf{S}_m^{*(j)}), \boldsymbol{\theta}^{(j)}$ we simulate $z(\mathbf{s}_0^{(j)})$ from the normal distribution with mean $\boldsymbol{x}(\mathbf{s}_0)^T\boldsymbol{\beta}^{(j)} + \tilde{w}^{(j)}(\mathbf{s}_0)$ and variance $\sigma_\epsilon^{2(j)}$ for $j = 1, \ldots, J$ where J is the total number of MCMC simulation. Finally, we form ergodic averages of $z(\mathbf{s}_0^{(j)})$, or its transformed values, to estimate features of the posterior predictive distribution.

We perform model validation using the root mean square prediction error (RMSPE) and also a cross-validation version of it. We also report the mean absolute prediction error to confirm robustness of the findings. The cross-validation RMSPE is calculated by first setting aside and then predicting each of the n observations in turn and then calculating $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(z(\mathbf{s}_i) - \hat{z}(\mathbf{s}_i))^2}$ where $\hat{z}(\mathbf{s}_i)$ is the cross-validation prediction for the observation $z(\mathbf{s}_i)$. We also use the Bayesian predictive model choice criterion (PMCC) proposed by Gelfand and Ghosh (1998) using independent predictive replicate $Z_{\text{rep},i}$ at location $\mathbf{s}_i$ of the observed data. The PMCC is sum of two parts: a goodness of fit G and penalty P where $G = \sum_{i=1}^{n}(z(\mathbf{s}_i) - E(Z_{\text{rep},i}))^2$ and $P = \sum_{i=1}^{n} \text{Var}(Z_{\text{rep},i})$.

4 Practical examples

4.1 Scallop data example

We consider the scallop data example studied by Ecker and Gelfand (1999) to illustrate the fitting and performance of the proposed anisotropic models. In this data set, recorded is the number of scallop catches for the year 1990 from 146 different locations in the North Atlantic near the City of New York. Following Ecker and Gelfand we also log-transform the data to reduce variability and to encourage Gaussianity. There is no spatially varying covariate available for this data set. Hence we work with a constant mean surface taking the value β at any location $\mathbf{s}$ within the study region. The mean parameter β is given the normal prior distribution with mean 0 and variance 10^4. We assume the exponential correlation function for the underlying GP and following Ecker and Gelfand we assume the uniform prior distribution $U(0.001, 30)$ for the decay parameter ϕ. This prior distribution allows for an effective range between 0.1 to 3000 kilometres. As mentioned before, $1/\sigma_\epsilon^2$ and $1/\sigma_w^2$ are assigned the Gamma prior distribution with parameters 2 and 1.

We have discussed several anisotropic models corresponding to different choice of knots and compared these next using out of sample cross-validation methods. To facilitate model comparison we split the data set into a training set, with 136 observations, and a validation set with the remaining 10 observations. The validation observations have been chosen to be at the same 10 sites as in the Ecker and Gelfand paper so that we can make a fair comparison of out of sample predictive performances of their model with that of the proposed ones.

We first select the number of knot locations using the RMSPE and the mean absolute prediction error (MAPE) based on the 10 validation observations. Table 1

provides the validation error estimates for the knot sizes of 100, 136, 225 and 400 for the two models where the knots are chosen by a space-filling design and the other one using a random placement of the knots. We include the model with a knot size of 136 since there are 136 observations in the fitting data set. As expected, the error estimates first decrease with the increasing knot-size and then start to increase after reaching a plateau of minimum values. The model with 225 knots and a random space-filling design seems to be the best according to the two error criteria. Henceforth, we proceed with this model, denoted by RSF225.

We now compare the best performing model with the following relevant modelling suggestions. We compare the performances of simple Kriging, and two models compared by Ecker and Gelfand: one with a general exponential covariance structure but with anisotropy as defined by their Equation (10) and the other with six parameter range anisotropic Matérn family as defined by their Equation (13). We denote these models by EGM I and II respectively. Table 2 provides the validation error estimates. The model RSF225 performs the best compared to all the other models including the EGM I model and the default GPP model with 136 knots.

Finally, we examine the predictions made using the anisotropic EGM I and the RSF225 at the 10 validation sites in Table 3. The values for the EGM II model are taken from the Ecker and Gelfand paper. The prediction standard deviations are lower for the proposed RSF225 and also as expected, the individual predictions are closer to the actual observed values.

4.2 London air pollution data example

As a second data example we consider the annual NO_2 air pollution data from 91 monitoring sites within Greater London and its suburb for the year 2011. In addition to these monitored data, we also make use of output of a numerical model, Air Quality Unified Model (AQUM) developed by Savage *et al.* (2013). AQUM is a large computer simulation model and uses emission inventory and many meteorological variables such as wind speed and direction to produce air pollution estimates at 1-kilometre square grids. We use the AQUM outputs in a downscaler regression model following Sahu *et al.* (2009) and Berrocal *et al.* (2010). Throughout, we model the data on the square root scale that encourages symmetry and normality. However, all the predictions are performed and compared on the original scale of the data for ease of interpretation.

We adopt model (6) with three covariates: AQUM values on the square-root scale, a rural-urban indicator and a roadside indicator as has been detailed in Lee *et al.* (2015). Here we compare four modelling methods as follows. The first method is the full spatial random effects model with the exponential covariance function denoted by

GEM16. We compare this base model with the following models: FCL16 for which 16 knots are clustered within a smaller rectangular sub-region covering the city of Westminster, FSF16 where 16 knots are selected according to a space filling design and kept fixed, and finally we consider the random space filling designs for knot selection with 9, 16 and 25 knots denoted respectively by RSF9, RSF16 and RSF25. All of the models are implemented with the exponential covariance function. The results reported below, however, remain unchanged qualitatively if the Matérn model is assumed instead. We have set the decay parameter ϕ to a fixed value 0.02 since this produced the best predictive performance for the models. In general, a tuning experimentation is required to choose the decay parameter value.

The first part of Table 4 shows the values of the PMCC (Gelfand and Ghosh, 1998). According to PMCC, we see that the RSF16 and RSF25 model are the best, although it has a higher G term than FSF16. The random placing of the knots is able to reduce the predictive penalty P term substantially but at the cost of increasing the G term. This however is a not a concern since the out of sample predictions as summarised by the cross-validation RMSPE and MAPE. The model RSF16 reduces the RMSPE's for the GEM16 model by about 75% pointing to a substantial gain. Figure 6 provides an interpolated surface showing the posterior probability of the knot-locations for the RSF25 model. The plot reveals that locations closer to the observation sites are slightly more likely to be selected as knots. An interpolated prediction surface, along with the observed values of NO_2 concentrations for a selection of sites to enhance readability, is shown in Figure 7. The map shows very good agreement between the predictions and observations.

5 Discussion

This paper finds that the GPP models, which originated as dimension reduction methods, are also able to generate flexible non-stationary and anisotropic models for spatial data. The paper demonstrates that structured selection of the knots leads to structured form of non-stationary and anisotropic models. The paper investigates the nature of anisotropy generated by these models and shows that the models generate different general forms of anisotropy which can accommodate the known types such as geometric and zonal anisotropy, see e.g. Chiles and Delfiner (2012) and Zimmerman (1993). These models are also shown to perform well using out of sample cross-validation predictions for two practical examples.

Novelty of our proposal also lies in recommendation of the models even for a smaller number of data points where dimension reduction is not required. Theoreti-

cal investigation and empirical evidence from two practical examples confirm that a random space filling design for knot selection for the predictive processes is the best which has been also observed by Guhaniyogi *et al.* (2011), but for large data sets in the context of dimension reduction. In this article our focus has been on using the GPP method even for smaller data sets. The predictive processes will also work for large data sets, but then the number of knot locations must be taken to be much smaller than the number of observations.

Future work will explore these methods in spatio-temporal data modelling settings with the added complexity of dynamic knot-designs at different time points. Work here will extend the space-time GPP models implemented in Bakar and Sahu (2015). Extension is also required for multivariate spatial data modelling.

Acknowledgements

The authors gratefully acknowledge the UK Met Office for providing the AQUM data.

References

Bakar, K. S. and Sahu, S. K. (2015). `spTimer`: Spatio-temporal bayesian modelling using `r`. *Journal of Statistical Software*, **63**(15), in press.

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

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, Boca Raton, 2nd edition.

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

Bolin, D. and Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Annals of Applied Statistics*, **5**(1), 523–550.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the spde approach. *Advances in Statistical Analysis*, **97**(2), 109–131.

Chiles, J.-P. and Delfiner, P. (2012). *Geostatistics*. John Wiley and Sons, Hoboken, 2nd edition.

Cressie, N. A. C. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons, New York.

Ecker, M. and Gelfand, A. (1999). Bayesian modeling and inference for geometrically anisotropic spatial data. *Mathematical Geology*, **31**, 67–83.

Finley, A. O., Banerjee, S., and Gelfand, A. E. (2015). `spBayes` for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, **63**(13), 1–28.

Fuglstad, G.-A., Simpson, D. P., Lindgren, F. K., and Rue, H. (2015). Exploring a new class of non-stationary spatial gaussian random fields with varying local anisotropy. *Statistica Sinica*, **25**(1), 115–133.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

Gelfand, A. E., Sahu, S. K., and Holland, D. M. (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics*, **23**, 565–578.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). Efficient metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith., editors, *Bayesian Statistics 5*, pages 599–607. Oxford University Press.

Guhaniyogi, R., Finley, A. O., Banerjee, S., and Gelfand, A. E. (2011). Adaptive gaussian predictive process models for large spatial datasets. *Environmetrics*, **22**(8), 997–1007.

Guhaniyogi, R., Finley, A. O., Banerjee, S., and Kobe, R. K. (2013). Modeling complex spatial dependencies: Low-rank spatially varying cross-covariances with application to soil nutrient data. *Journal of Agricultural Biological and Environmental Statistics*, **18**(3), 274–298.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, **5**(2), 173–190.

Ingebrigtsen, R., Lindgren, F. K., and Steinsland, I. (2014). Spatial models with explanatory variables in the dependence structure. *Spatial Statistics*, **5**(1), 20.

Katzfuss, M. (2013). Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics*, **24**(3), 189–200. 10.1002/env.2200.

Konomi, B. A., Sang, H., and Mallick, B. K. (2014). Adaptive bayesian nonstationary modeling for large spatial datasets using covariance approximations. *Journal of Computational and Graphical Statistics*, **23**(3), 802–829.

Lee, D. P., Mukhopadhyay, S., Rushworth, A., and Sahu, S. K. (2015). A rigorous statistical framework for estimating the long-term health impact of air pollution, with application to respiratory hospitalisation risk in england between 2007 and 2011. Technical report, University of Glasgow.

Lindgren, F., Rue, H., and Lindstrm, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(4), 423–498.

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

Sahu, S. K., Yip, S., and Holland, D. M. (2009). Improved space-time forecasting of next day ozone concentrations in the eastern u.s. *Atmospheric Environment*, **43**, 494–501.

Sampson, P. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covarance structure. *Journal of the American Statistical Association*, **87**(417), 108–119.

Savage, N. H., Agnew, P., Davis, L. S., Ordóñez, C., Thorpe, R., Johnson, C. E., O'Connor, F. M., and Dalvi, M. (2013). Air quality modelling using the met office unified model (aqum os24-26): model description and initial evaluation. *Geoscientific Model Development*, **6**(2), 353–372.

Schmidt, A. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society, Series B*, **65**(3), 743–758.

Spiegelhalter, S. D., Best, N. G., Carlin, B. P., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, **64**(4), 583–639.

Zimmerman, D. L. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, **25**(4), 453–470.

	100		136		225		400	
	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE
SF	0.85	0.74	0.87	0.74	0.88	0.74	0.80	0.66
RSF	0.78	0.67	0.77	0.66	0.73	0.62	0.79	0.68

Table 1: Validation error estimates for the two models: space filling (SF) and random space filling (RSF) with different number of knots.

Kriging		EGM I		EGM II		RSF225	
RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE
1.08	0.63	0.99	0.84	0.89	0.77	0.73	0.62

Table 2: Validation error estimates for different models for the 1990 scallop data set.

		EGM I		RSF225	
Site	$Z(s)$	Mean	SD	Mean	SD
1	1.946	2.181	1.331	2.00	1.141
2	1.792	2.745	1.372	2.558	1.283
3	4.007	3.666	1.369	3.455	1.286
4	4.331	4.318	1.325	4.370	1.302
5	5.501	4.463	1.330	4.754	1.235
6	5.645	4.456	1.309	4.358	1.295
7	5.620	4.131	1.369	4.780	1.112
8	4.394	3.718	1.374	3.525	1.252
9	3.332	2.756	1.240	3.10	1.263
10	0	1.216	1.304	0.797	1.312

Table 3: The predicted values along with their standard deviations using the two models.

Model	GEM16	FCL16	FSF16	RSF9	RSF16	RSF25
G	255	155	154	155	155	155
P	5847	529	524	12	11	11
G+P	6102	684	678	167	166	166
RMSPE	85.37	21.48	21.40	21.48	21.25	21.46
MAPE	76.82	16.87	16.70	17.09	16.74	16.87

Table 4: PMCC values and summaries of leave one out cross-validation values for models fitted to annual NO_2 data from 91 monitoring sites in London for the year 2011. GEM16 stands for the full dimensional spatial random effects model. FCL16 is the model based on 16 clustered knots and FSF16 is the model with 16 fixed knots chosen according to a space filling design. RSF9, RSF16 and RSF25 denotes models with random space filling design with 9, 16 and 25 knots respectively.

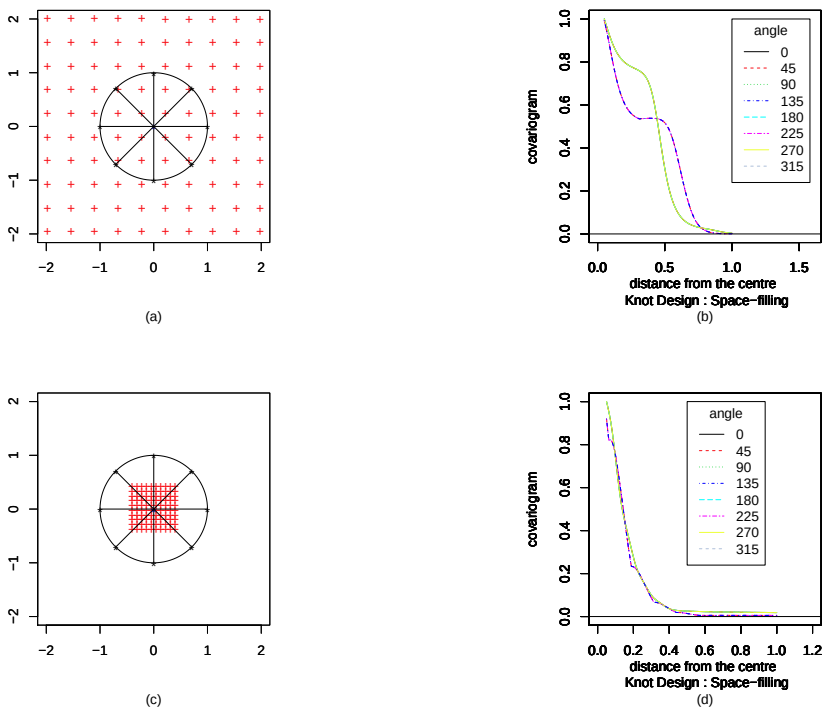

Figure 1: Two fixed knot-designs (panels (a) and (c)) and the corresponding implied covariance function plots (panels (b) and (d)) when distances are calculated from the origin at different angles, which are shown as black straight line segments in panels (a) and (d).

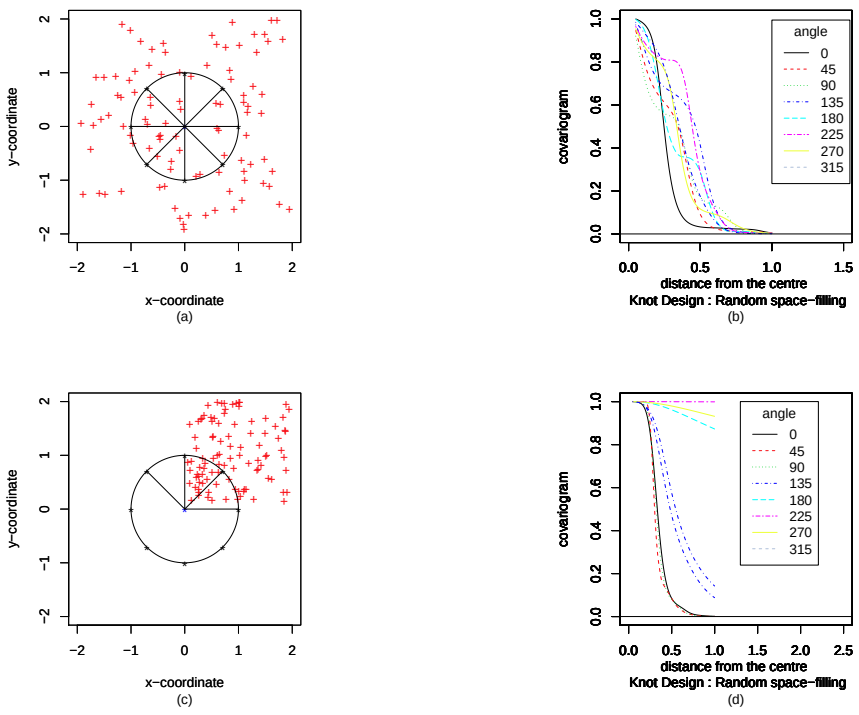

Figure 2: Two random knot-designs and the implied covariance function plots when distances are calculated from the origin at different angles. The knots in panel (a) are chosen at random locations over the whole study region and the knots in panel (d) are randomly chosen to lie in the first quadrant only.

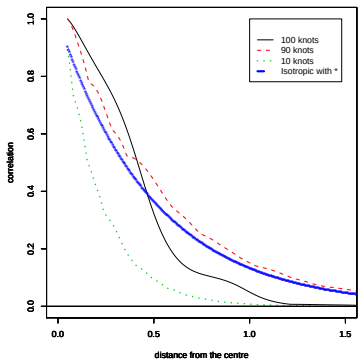

Figure 3: A plot of the correlation functions for different knot sizes when the knots are placed randomly and the distances are calculated along the perpendicular line. The plot for the isotropic correlation function is super imposed for comparison purposes. Other correlation parameters are taken to be the same as in panel (b) of Figure 2.

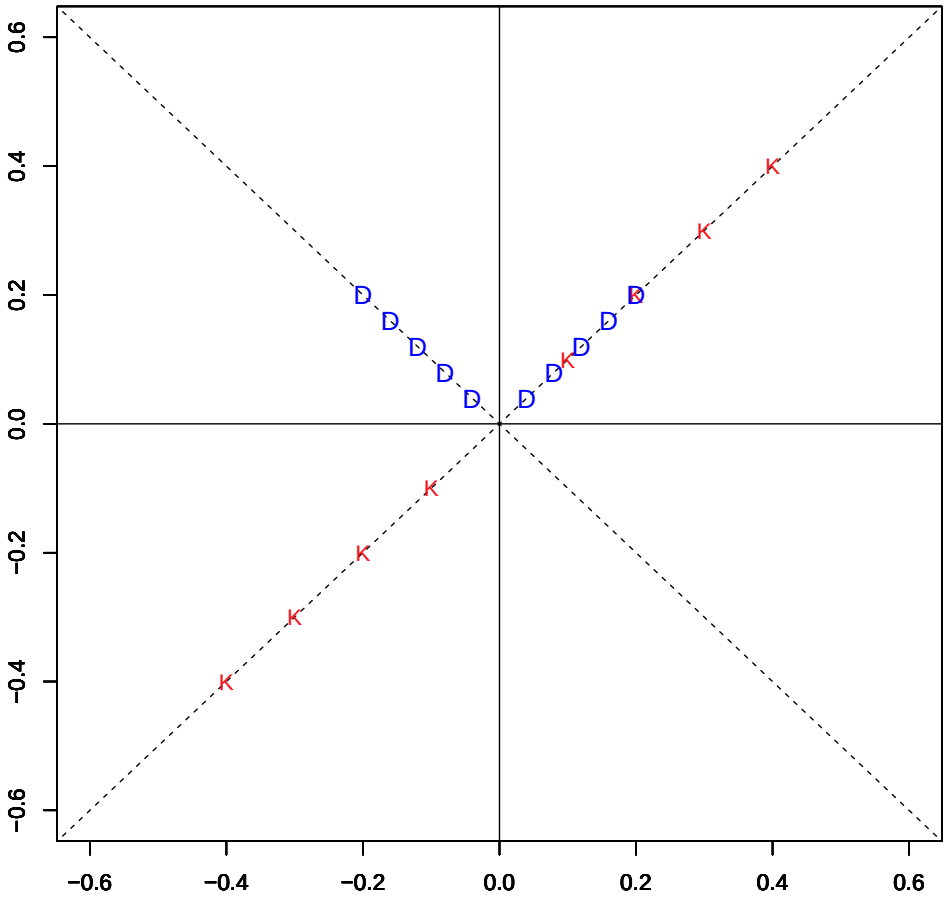

Figure 4: Illustration of the arrangement of knots, denoted by K and data locations of interests, denoted by D.

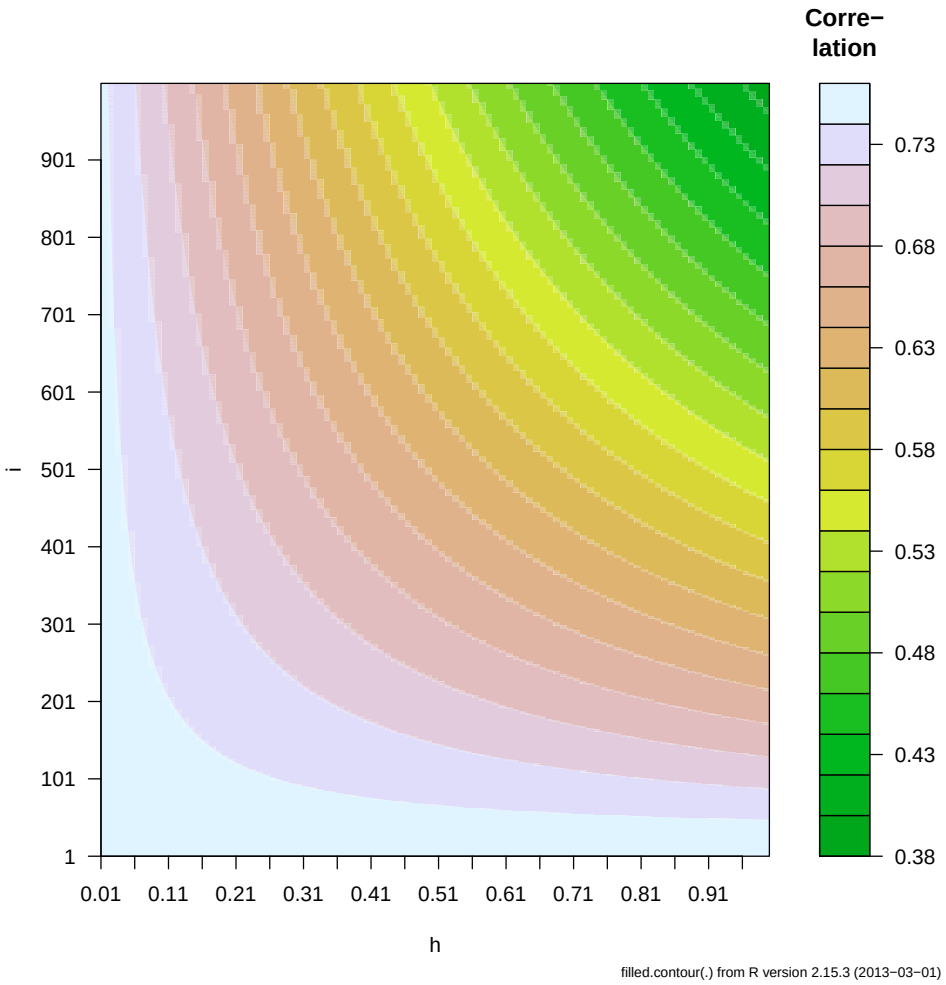

Figure 5: Surface plot of correlation values between the observations at the origin and a location determined by i, $i = 1, \ldots, 1000$, for various values of h, which determines the knot locations.

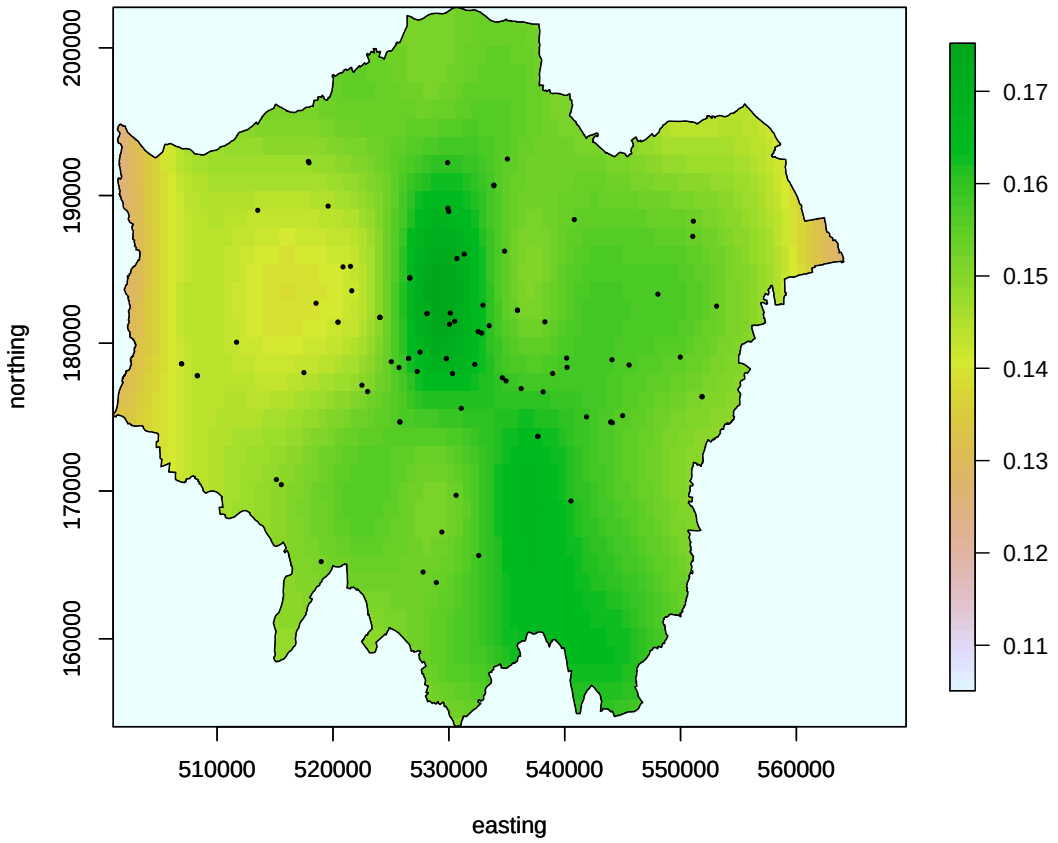

Figure 6: An interpolated surface the posterior probability of the knot-locations. The observation sites are superimposed except for the 7 seven sites which fall outside the Greater London boundary shown in the map.

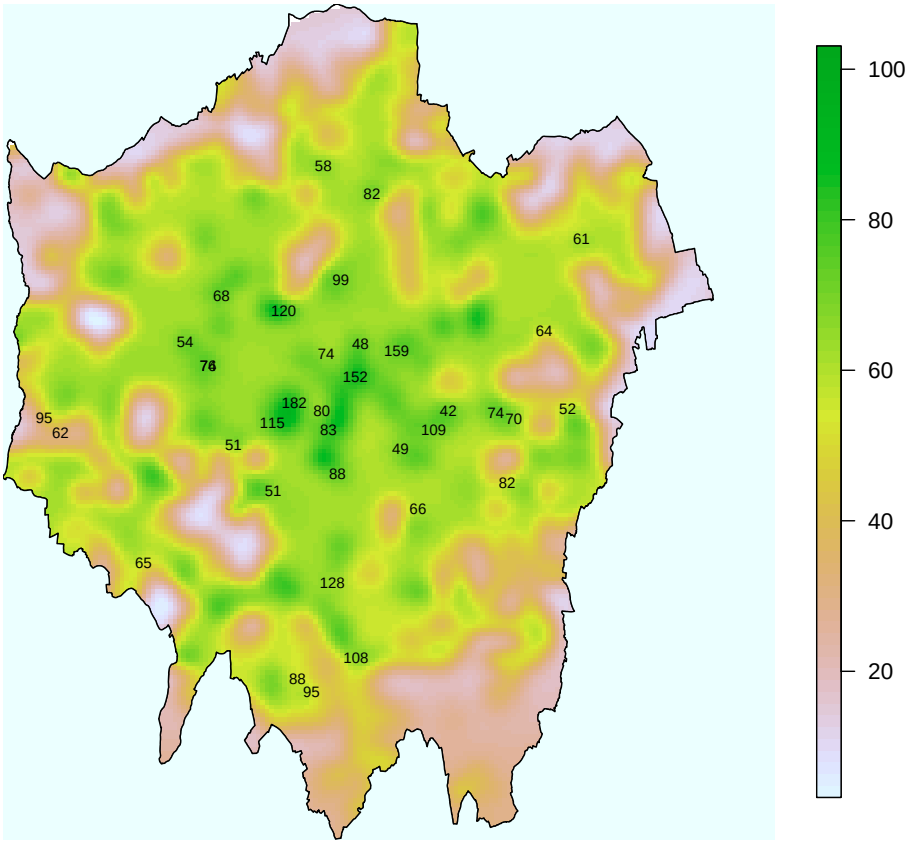

Figure 7: Prediction surface showing the annual average NO_2 values. Actual observed values from selected sites are superimposed. Values from the remaining sites are omitted to enhance readability.