On generating a flexible class of anisotropic spatial models using Gaussian predictive processes

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

February 24, 2015

Abstract

This article proposes a flexible class of non-stationary and anisotropic 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 sizes and studying the nature of anisotropy 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 anisotropic covariance functions not yet studied in the literature. These new covariance functions give rise to new flexible and accurate Bayesian predictive models but do not complicate the fitting and analysis methods unlike other models based on anisotropic covariance functions. 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.

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

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

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. Popularity of these models also come from their abilities to provide inference and prediction capabilities on dis-aggregated spatial scales. This advance is also fuelled by increasing computing power and the development of modern computational techniques based on Markov chain Monte Carlo (MCMC) and Integrated Nested Laplace Approximation (INLA) and others implemented as 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).

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. However, such simplicity does not often represent reality and practical data sets exhibit non-stationarity and anisotropy. Spatial non-stationarity arises when features of the data distribution, e.g., means and variances, depend on the actual spatial locations where those are observed. Anisotropy, on the other hand, points to un-equal and un-even variation of the data as a function of either the direction or the distance between locations where the data points have been collected. These un-even features are hard to generalise into a simplistic parametric model and hence there has been a noticeable slow progress in the literature regarding the use of non-stationary and anisotropic spatial models in place of their stationary and isotropic counter parts. However, 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 nonstationary 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.

The main objective of this paper is to introduce a method to generate a flexible class of anisotropic, as well as non-stationary, spatial models which are also based on Gaussian processes. The generating mechanism relies on specification of a Gaussian predictive process (GPP), see Banerjee *et al.* (2008) and flexibility is added through the choice of a set of knot-locations where the predictive process is anchored. A GPP is simply defined as the process induced by Kriging. Given a set of m 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 realisations at the knots denoted by $\mathbf{w}^* = (w(\mathbf{s}_1^*), \ldots, w(\mathbf{s}_m^*))^T$. In particular,

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

There are several questions regarding the use of $\tilde{w}(\mathbf{s})$ in spatial modelling. What are the covariance properties of this new process $\tilde{w}(\mathbf{s})$? How do those properties change as the *knots* are moved around in the study region D? 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 m influence the covariance structure of $\tilde{w}(\mathbf{s})$? What effect does the assumed covariance structure of the GP $w(\mathbf{s})$ have on $\tilde{w}(\mathbf{s})$? It is well known that $\tilde{w}(\mathbf{s})$ *a*lways generates a non-stationary and anisotropic spatial process (Banerjee *et al.*, 2008). However, there is no comprehensive study in the literature which is able to answer all these questions.

Our main contribution here is to investigate these issues in detail with both theoretical and practical examples in order to develop accurate predictive models. Here we find that $\tilde{w}(\mathbf{s})$ defines a new class of flexible spatial models which are able to capture highly un-structured concepts of anisotropy as yet un-explored in the literature. This new concept of anisotropy produces explicit non-stationary processes based on parametric covariance functions and it can lead to many traditional notions of anisotropy: geometric, sill, nugget, range and zonal, see e.g. Zimmerman (1993); Ecker and Gelfand (1999). We compare the new anisotropic covariance models with the default Gaussian process (GP) models in illustrative theoretical examples and also show better model fitting and prediction results in two real life practical examples. We also argue that that a denseness condition on the set of knot locations is required to guarantee a rich class of anisotropic models.

To illustrate the main ideas of this paper consider the real line $\mathbb{R}^1$ to be the study region D. With one knot point, 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 (1) where $w^*(0) \sim N(0,1)$. It is now easy to see that 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. It is also clear that the covariance function will also be non-monotonic with respect to the distance $|s - s'|$, even though the underlying stochastic process is Gaussian with a monotonic covariance function. Further complexity in the covariance function are easily introduced

by: (i) having more than one knot-locations so that $\tilde{w}(s)$ is based on $\mathbf{w}^*$ where $m > 1$, (ii) assuming the knot locations to be assigned at random over a finite sub-interval in D (iii) assuming specific clustering processes for the knot-locations. Clearly, using this process we are able to construct a model possessing rich anisotropic and non-stationary covariance structures. Further, illustrations with specific examples are provided in Section 2.

Traditionally, predictive processes have been used as dimension reduction techniques in the following sense. In the full dimensional model, each observation of a spatial random variable, say $Z(\mathbf{s})$, is assigned a spatial random effect $w(\mathbf{s})$ which is often assumed to follow a GP as above. However, when the number of observations n is very large, it is prohibitive to work with the full dimensional process $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T$, giving rise to the so called *big-n problem*, since working in a Bayesian inference set-up requires inversion of $n \times n$ matrices in iterative MCMC based model fitting and prediction algorithms. The dimension reduction technique, proposed by Banerjee *et al.* (2008), instead allows working with the m dimensional $\mathbf{w}^*$ where m is much smaller than n and here, each $w(\mathbf{s}_i)$ is replaced by the conditional expectation $\tilde{w}(\mathbf{s}_i)$ as in (1). This approximation makes model fitting and prediction tasks much easier even when n is very large and consequently Bayesian `R` packages have been developed, e.g. `spBayes` (Finley *et al.*, 2007) and `spTimer` (Bakar and Sahu, 2015). In both these packages it is recommended that the required knot-locations and their number m are specified either by trial and error on the basis of out of sample predictions or by using space filling designs. However, Guhaniyogi *et al.* (2011) propose fully adaptive GPP models based on random knot locations which are distributed according to a point pattern model for large data sets.

Our proposal here is to use the predictive process $\tilde{w}(\mathbf{s})$ even for small sizes where dimension reduction is not needed. We propose to make m larger than n since this allows us to gain important model flexibility regarding non-stationarity and anisotropy, as illustrated with two practical examples in Section 5. As in Guhaniyogi *et al.* (2011), we also propose random knot specification, and use their estimation method that achieves better model fitting and prediction results. The required computation burden, however, is still kept manageable as long as m is not allowed to be so large as to hit the big-n problem, described above, for which dimension reduction through a predictive process has been proposed in the first place.

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 3 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 4. Section 5 contains illustrations of the methods using two practical examples: one for the scallop catch data set and the other modelling PM_{10} pollution levels in the City of London in 2011. A few summary remarks are placed in Section 6.

2 GPP method for generating anisotropic models

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. To solve the big-n problem, GPP approximation replaces $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))'$ by

$$\tilde{\mathbf{w}} = E\left[\mathbf{w}|\mathbf{w}^*\right]$$

where $(\mathbf{w}, \mathbf{w}^*)$ is assumed to follow the multivariate Gaussian distribution with mean zero and a covariance matrix constructed using the underlying GP. Thus, marginally $\mathbf{w}^*$ is specified as $N(\mathbf{0}, \Sigma_{w^*})$ where $\Sigma_{w^*} = \sigma_w^2 S_{w^*}$ where S_{w^*} is an $m \times m$ correlation matrix whose entries are formed using the correlation function $C(\cdot)$ of the underlying GP. Multivariate Gaussian theory yields that

$$\tilde{\mathbf{w}} = C^* S_{w^*}^{-1} \mathbf{w}^* \tag{2}$$

where C^* is the $n \times m$ cross-correlation matrix between $\mathbf{w}$ and $\mathbf{w}^*$, i.e. $(C^*)_{ij} = C(|\mathbf{s}_i - \mathbf{s}_j^*|)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The spatial random effects surface, $\tilde{\mathbf{w}}$, is now based on linear functions of the m-dimensional $\mathbf{w}^*$ instead of the n-dimensional $\mathbf{w}$. This leads to a reduction of computational burden when m is much smaller than n. The focus of this paper, however, is not on the dimension reduction aspects, but on studying the implied covariance structure of $\tilde{w}(\mathbf{s})$.

For two locations $\mathbf{s}$ and $\mathbf{s} + \mathbf{h}$, separated by the vector $\mathbf{h}$, the correlation between $\tilde{w}(\mathbf{s})$ and $\tilde{w}(\mathbf{s} + \mathbf{h})$ is given by:

$$\text{Cor}\left(\tilde{w}(\mathbf{s}), \tilde{w}(\mathbf{s} + \mathbf{h})\right) \equiv \tilde{C}(\mathbf{s}, \mathbf{s} + \mathbf{h}) = \mathbf{c}^*(\mathbf{s})^T S_{w^*}^{-1} \mathbf{c}^*(\mathbf{s} + \mathbf{h})$$

where $\mathbf{c}^*(\mathbf{s})$ denotes the $m \times 1$ correlation vector between $w(\mathbf{s})$ and $\mathbf{w}^*$, given by $(C(|\mathbf{s} - \mathbf{s}_1^*|), \ldots, C(|\mathbf{s} - \mathbf{s}_m^*|))^T$. Clearly, the $\tilde{C}(\mathbf{s}, \mathbf{h})$ depends on both $\mathbf{s}$ and $\mathbf{h}$ and not only on the separation vector $\mathbf{h}$ or the distance $|\mathbf{h}|$. As a result, the model specification with $\tilde{w}(\mathbf{s})$ as the spatial effects will also imply non-stationary and anisotropic

correlation structure. Further exploration of the correlation structure is developed using the traditional semivariogram, denoted by, $\tilde{\gamma}(\mathbf{s}, \mathbf{h})$ of the spatial effects $\tilde{w}(\mathbf{s})$ as follows:

$$\begin{aligned} 2\tilde{\gamma}(\mathbf{s}, \mathbf{h}) &= \mathrm{Var}\left[\tilde{w}(\mathbf{s}) - \tilde{w}(\mathbf{s}+\mathbf{h})\right] \\ &= E\left[\tilde{w}(\mathbf{s}) - \tilde{w}(\mathbf{s}+\mathbf{h})\right]^2 \\ &= E\left[\mathbf{c}^*(\mathbf{s})^T S_{w^*}^{-1}\mathbf{w}^* - \mathbf{c}^*(\mathbf{s}+\mathbf{h})^T S_{w^*}^{-1}\mathbf{w}^*\right]^2 \\ &= E\left[\left(\mathbf{c}^*(\mathbf{s}) - \mathbf{c}^*(\mathbf{s}+\mathbf{h})\right)^T S_{w^*}^{-1}\mathbf{w}^*\right]^2 \\ &= \left(\mathbf{c}^*(\mathbf{s}) - \mathbf{c}^*(\mathbf{s}+\mathbf{h})\right)^T S_{w^*}^{-1}\left(\mathbf{c}^*(\mathbf{s}) - \mathbf{c}^*(\mathbf{s}+\mathbf{h})\right). \end{aligned}$$

Non-stationarity of the $\tilde{w}(\mathbf{s})$ is apparent from the dependence of the semivariogram, $\tilde{\gamma}(\mathbf{s}, \mathbf{h})$ on both $\mathbf{s}$ and $\mathbf{h}$. Traditional variogram analysis proceeds for stationary models where $\tilde{\gamma}(\mathbf{s}, \mathbf{h})$ is assumed to depend on the separation vector $\mathbf{h}$ only. Here it depends not only on $\mathbf{s}$ and $\mathbf{h}$ but also implicitly on the chosen knot-locations $\mathbf{S}_m^*$ and m, the number of knots. Moreover, the parameters of the correlation function of the Gaussian process will also influence $\tilde{\gamma}(\mathbf{s}, \mathbf{h})$.

Can we treat $\tilde{\gamma}(\mathbf{s}, \mathbf{h})$ as a legitimate semivariogram? The answer must be no since $\tilde{\gamma}(\mathbf{s}, \mathbf{h})$ is no longer an even function of $\mathbf{h}$ since $\tilde{\gamma}(\mathbf{s}, \mathbf{h}) \neq \tilde{\gamma}(\mathbf{s}, -\mathbf{h})$ due to its dependence on $\mathbf{s}$ and also on the knots $\mathbf{S}_m^*$. However, we propose to treat this as a function of the separation distance $|\mathbf{h}|$ and study its properties for varying configurations of $\mathbf{s}$, $\mathbf{h}$ and $\mathbf{S}_m^*$. In the sequel we also compare this with the isotropic correlation function of the underlying stationary GP to investigate the nature of anisotropy and non-stationarity introduced by the GPP specification.

The dependence of $\tilde{\gamma}(\mathbf{s}, \mathbf{h})$ on both $\mathbf{s}$ and $\mathbf{h}$ makes the comparison of the semivariograms for different distances or knot specifications difficult. We deal with this difficulty by first fixing a 'central' location, $\mathbf{s}^{**}$, which is assumed to be the centroid of the study region D and then calculating distance $|\mathbf{h}|$ between $\mathbf{s}^{**}$ and $\mathbf{s}^{**} + \mathbf{h}$ in different directions within D. This allows us to study the induced directional correlation structure as data locations move away from the centre $\mathbf{s}^{**}$ to the edges of D. If instead, the intention is to study correlation structure across diagonal distances within D, then also we can use the same method but then we have many choices of the particular point $\mathbf{s}^{**}$ from which to calculate distance.

The number, m, and configuration of the knots plays a major role in dictating the nature of anisotropy as the examples below will illustrate. A novel proposal here is to allow the knot-locations to be random. 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 covariance function $\tilde{\gamma}(\mathbf{s}, \mathbf{h})$ will be a random quantity if the knot locations $\mathbf{S}_m^*$ are also random. The true covariance function will then be the expected values of

these random covariance function where expectation is taken over the distributions of m and $\mathbf{S}_m^*$ denoted respectively by $\pi(m)$ and $\pi(\mathbf{S}_m^*)$. In other words,

$$2\tilde{\gamma}(\mathbf{s},\mathbf{h}) = E_{\pi(m),\pi(\mathbf{S}_m^*)} E\left[\{\tilde{w}(\mathbf{s}) - \tilde{w}(\mathbf{s}+\mathbf{h})\}^2 | m, \mathbf{S}_m^*\right]$$

where we continue to use $\tilde{\gamma}(\mathbf{s},\mathbf{h})$ to denote the expected semivariogram. We use Monte Carlo sampling to evaluate the outer expectation 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 evaluate the inner expectation $E\left[\{\tilde{w}(\mathbf{s}) - \tilde{w}(\mathbf{s}+\mathbf{h})\}^2 | m_l, \mathbf{S}_{m_l}^*\right]$. Finally, we approximate $\tilde{\gamma}(\mathbf{s},\mathbf{h})$ by

$$\frac{1}{2L}\sum_{\ell=1}^{L} E\left[\{\tilde{w}(\mathbf{s}) - \tilde{w}(\mathbf{s}+\mathbf{h})\}^2 | m_l, \mathbf{S}_{m_l}^*\right].$$

Thus the most general method we propose to generate stochastic processes having non-stationary and anisotropic correlation structure 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 theoretical examples one each on one and two dimensional sub-spaces.

3 Theoretical examples

Assume that the study region D is the compact region, $[-1,1]$ and $[-1,1]\times[-1,1]$ in one and two dimensional spaces respectively, so that the centroid $\mathbf{s}^{**}$ in both cases is the origin. We consider the following four scenarios of knot specification. The first design is space filling with $m = 25$ knots within D and the second design corresponds to random placement of the knots according to the complete spatial randomness (CSR) point pattern within D. The third scenario places all the knots within a particular sub-region of D which we choose to be the central region $[-0.25, 0.25]$ in one dimension and $[-0.25, 0.25]\times[-0.25, 0.25]$ in two dimension in our illustration. This third design allows us to see the effect of clustering of the knots in one particular sub-region which may have been incorrectly chosen by the modeller. The final design considered here is also a CSR but we take the number of knots m to be selected at random from the integers 1 to 25. This allows us to investigate any possible gain in terms of extra flexibility in covariance structure by having a random number of knots. For the examples in this section, we assume exponential correlation function with ϕ fixed at 0.2 for illustrations.

Figure 1 illustrates the theoretical semivariograms under four different knot designs for the one dimensional example. In each sub-panel we plot two semivariograms against values of distance from the origin: one in the positive and the other in the negative direction. As expected the two directional semivariograms collapses into one for the two fixed knot designs in the first column of the plot. The two semivariograms vary slightly in the second column due to the random placing of the knots, but essentially the same phenomenon occurs because of the averaging used to evaluate the expectation over the distribution of the knots. The predominant wave pattern in the two semivariograms in the first column emerges from the proximity of the locations, for which distance has been calculated from the origin s^{**}, to the nearest knot-location. The semivariogram has local maxima when $s^{**} + h$ coincides with a knot location s^* since there is a random variable $w(s^*)$ with its full variance instead of the conditional expectation, with reduced variation, for locations away from the knots. Clustering of the knots has the effect of creating anisotropic semivariograms when the two locations are within the region where the knots are clustered. The semivariograms smooth out when at least one of the locations are away from the clustering region. Lastly, we note the similarity between the semivariograms in the two panels in the second column. Here the effect of randomness in the number of knots, m, has been averaged out by the outer expectation with respect to $\pi(m)$. Thus, there is no extra flexibility to be gained by varying the number of knots.

Results from the second example are illustrated in Figure 2 where the semivariogram values are plotted against the radial distance for four different angles within the first quadrant. Only the first quadrant is considered here since the semivariograms will be same for the complimentary angles in the other quadrants. Distance is still calculated from the origin $\mathbf{s}^{**}$, but we consider four different angles for illustration purposes instead of the only two possible directions in the one dimensional example. In the first panel for the semivariogram at zero degrees, we see occurrence of local maxima when the point $\mathbf{s}^{**} + \mathbf{h}$ coincides with a knot location. The similar phenomenon is not observed for the semivariogram at other angles due to the non-coincidence of any of the $\mathbf{s}^{**} + \mathbf{h}$ with any knot-location for those angles. Figure 3 shows the positions of the knot-locations, the angular directions and the circles at fixed radii for which we illustrate the semivariograms. The effect of clustering of the knots can be seen in the bottom right panel of Figure 2 by omission of the local maxima beyond any distance from the last knot-location placed within the smaller sub-region. The semivariograms under the two random scatter designs for the knot-locations are similar and the conclusions here are also similar to the ones for the one dimensional example. The only difference to be noticed is that the semivariograms for different angles can

be somewhat different which confirms a form of angular anisotropy induced by these predictive processes.

The angular anisotropy, induced by the predictive processes, is studied further in Figure 4. Here each semivariogram in each sub-panel corresponds to a fixed radial distance r and the points for which the semivariogram values are calculated lie on the circumference of the circle. Given an angle, say ψ, between 0 and 2π, and a radius r, we consider the point $\mathbf{h} = (r\cos(\psi), r\sin(\psi))$ and then calculate the value of the semivariogram $\gamma(\mathbf{0}, \mathbf{h})$. For a fixed value of r, the semivariogram varies according to the angle ψ and shows angular anisotropy since the corresponding isotropic semivariogram will be a constant if r is fixed even if ψ varies. The semivariogram is essentially flat when the radius is small since there are not many locations placed within the circular region centred at the origin. Increasing the radius r implies increasing the number of knot-locations within the circular region and the variability increases as the effect in variation is greater with a greater number of nearby knot-points. Here the shape of the semivariograms are not always monotonic and hence the curves are for neither a geometric nor a zonal anisotropy.

How can this sort of unstructured flexible anisotropy can be guaranteed in the model? The plots in Figure 4 suggest that the semivariogram will have a local maxima if at least one of the two locations for which it is calculated coincides with a particular knot location. This in turn implies that a high level of anisotropy can be guaranteed if there is a knot-location present in 'every' neighbourhood containing any two possible locations in the study region D. Thus, to guarantee a degree of anisotropy we need to spread the knot-locations throughout the study region possibly using a space-filling design. Intuitively, we can now infer that a topological denseness condition on the set of knot-locations within the study region D will guarantee high level of anisotropy. Also, to encourage further flexibility we propose the use of random space filling designs where the data will help us to pin-point the optimal knot-locations.

4 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:

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

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

$$\pi(\mathbf{s}_j^*) = \frac{p(\mathbf{s}_j^*)}{\sum_{j=1}^{M} p(\mathbf{s}_j^*)} \tag{3}$$

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 $\tilde{w}(\mathbf{s}) = \mathbf{c}^*(\mathbf{s}) S_{w^*}^{-1} \mathbf{w}^*$ 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^2_ϵ and σ^2_w 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:

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

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2_\epsilon, \sigma^2_w, \nu, \phi)^T$ 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 (3) 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 $(\boldsymbol{\beta}, \sigma^2_\epsilon, \sigma^2_w, \nu, \phi)^T$. 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 $\mathbf{c}^*(\mathbf{s}_0)S^{-1}_{w^*}\mathbf{w}^*$, 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 $P = \sum_{i=1}^n \text{Var}(Z_{\text{rep},i})$.

5 Practical examples

5.1 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.

As in Ecker and Gelfand we explore the nature of anisotropy present in the scallop data by the semivariogram contour plots. These plots are based on distance between locations described by differences in both the coordinates in a two dimensional study region rather than the scalar distance. The empirical version, see Banerjee *et al.* (2015), is described briefly as follows. Assume that we have n observed values of the

spatial random variable $z(\mathbf{s}_i), i = 1, \ldots, n$ in the study region D which is a subset of R^2. First we calculate the separation distances h_x and h_y in the x and y coordinates respectively for each of the ${}^nC_2$ pairs of locations where to h_y is taken to be greater than zero to avoid the impact of the order of the pair of locations. These separation distances are aggregated into rectangular bins B_{ij} and the empirical semivariogram value for the (i, j)th bin is calculated by

$$\gamma^*_{ij} = \frac{1}{2N_{B_{ij}}} \sum_{\{(i,j):((k,l):(\mathbf{s}_k - \mathbf{s}_l) \in B_{ij}\}} (z(s_k) - z(s_l))^2 ,$$

where $N_{B_{ij}}$ is the number of sites in bin B_{ij}. The bin width in y axis is taken to be half of that for the x axis since $h_y > 0$ for all pairs but h_x's are unrestricted. Also the middle class of the x-axis is centred around zero since the differences are symmetric. The empirical semivariogram plot (ESC) is the three dimensional contour plot of γ^*_{ij} against the centre of the bin B_{ij} denoted (x_i, y_j). A smoothed version of this plot is called the empirical semivariogram contour plot (Banerjee *et al.*, 2015).

Analogous to the ESC plots we define theoretical semivariogram contour (TSC) plot as the smoothed contour plot of $E\left(\gamma^*_{ij}\right)$ against the centre (x_i, y_j) assuming that the expectation exists. Isotropic models will show circular contours while geometric anisotropy is captured by elliptical contours.

Figure 5 illustrates the TSC and ESC plots for the 1990 scallop data set. The top left panel of this figure, as expected, show the circular contours corresponding to a stationary Gaussian process. The top right panel is the ESC plot of the data where the plotted contours do not show any predominant circular or elliptical pattern. Rather it detects many local anisotropic patterns of different shapes pointing to the difficulty in modelling with isotropic and stationary models. The TSC plot for the 100-knot GPP model with a fixed space-filling design for the knot selection appears in the bottom left panel while the same for the random space filling design is depicted on the bottom right panel. Clearly, the patterns evident in the ESC plot of the data match much more closely with the patterns present in the TSC plots of the proposed models than those of an isotropic model, which conquers the quest of Ecker and Gelfand to search for possible anisotropic models for these data. All these anisotropic models are compared 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.

5.2 London air pollution data example

As a second data example we consider annual air pollution data from 16 monitoring sites within the City of Greater London. These monitoring sites along with the boundaries of the 32 local health authorities are shown in Figure 6. We illustrate with the particulate matter PM_{10} which are pollution particles less than 10 micrometers in size. 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.

Here we compare the following four methods for modelling the air pollution data. The first method is the full spatial random effects model with the Matérn covariance function denoted by GEM16. We compare this base model with the following models: FCL16 for which 16 knots are clustered within a smaller sub-region, 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. Figure 6 shows the smaller clustering sub-region and also the 16 knot-locations for the FSF16 model. All of the models are implemented with the Matérn covariance function where the smoothing parameter ν is kept fixed at 0.2 which encourages a lower level of smoothing than the exponential model at $\nu = 0.5$. The results reported below, however, remain unchanged qualitatively if the exponential model is assumed instead. We do not attempt to estimate ν in this small data example as there are problems in estimating all the parameters under a general Matérn model, see e.g. Zhang (2004). For the decay parameter ϕ we use a Gamma prior distribution with hyper-parameter values 1.06 and 1 having mean 1.06. These values achieved the best predictive performance for the models. In general, a tuning experimentation is required to choose the hyper-parameter values.

The first part of Table 4 shows the values of the PMCC (Gelfand and Ghosh, 1998). According to PMCC, we see that the RSF25 model is the best, although it has a higher G term. 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 RSF25 reduces the RMSPE's for the GEM16 model by about 45% pointing to a substantial gain. Table 6 examines these leave one out cross-validation predictions with further details providing the standard deviations of the predictions and the 95% prediction intervals. Exactly one out of the 16 prediction intervals fails to contain the true observed value. Otherwise, a broad agreement is seen between the observed and predicted values, but the model based predictions are generally smoother than the observations, as expected.

The parameter estimates for the two models, GEM16 and RSF25 models are provided in Table 5. The estimates for the intercept from both the models show large negative bias of the AQUM output as has been noted by Savage *et al.* (2013) and this negative bias is significant under the RSF25 model, which also shows that the slope is significant. This confirms that the AQUM model output has a significant positive relationship with the observed PM_{10} values even under a spatial model. In

fact, corresponding to the actual observations provided in Table 6, all the AQUM values (not shown here) were between 11 and 12. Both the models suggest strong spatial correlation with an effective range between 1 and 2 kilometres. The estimate of the spatial variance parameter (σ_w^2) is larger than that for the nugget effect σ_ϵ^2 under the RSF25 model which confirms that spatial variability is the dominant of the two variance components which in turn signifies presence of a spatial story. Lastly, Figure 7 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.

6 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. Theoretical 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. 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 would like to acknowledge the support of EPSRC grant EP/J017485/1 awarded to University of Southampton.

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 downcsaler for outputs from numerical models. *Journal of Agricultural, Biological and Environmental Statistics*, **15**, 176–197.

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 Carlin, B. P. (2007). `spBayes`: An `r` package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, **12**(4), 1–24.

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.

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.

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.

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.

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

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.

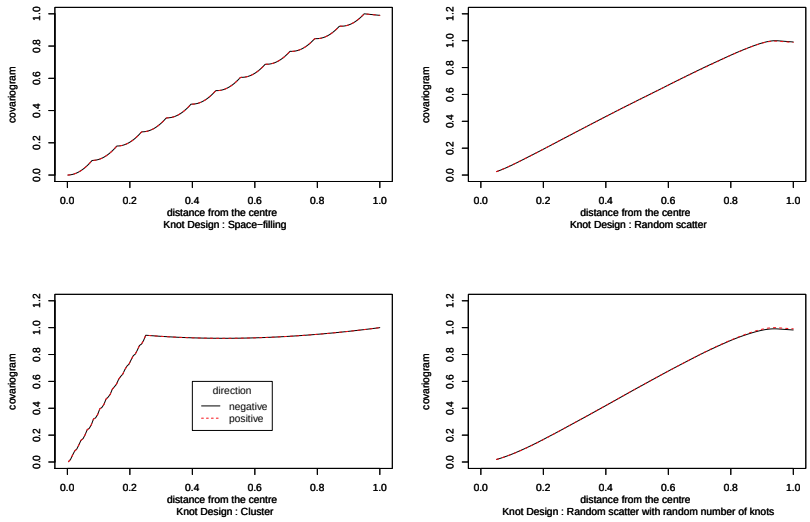

Figure 1: Semivariogram plots under different knot configurations. Solid line is used for the semivariogram in the positive direction from the origin and the dotted line is for the semivariogram in the negative direction.

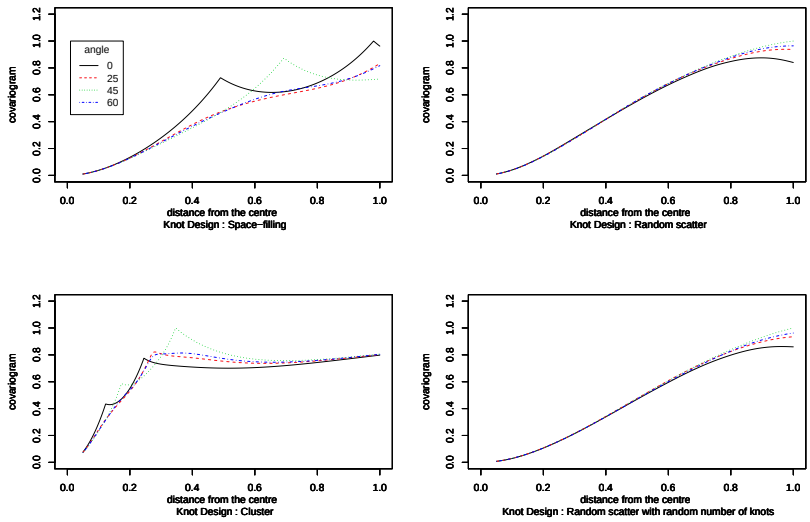

Figure 2: Semivariogram plots against radial distance for the two dimensional example under different knot configurations.

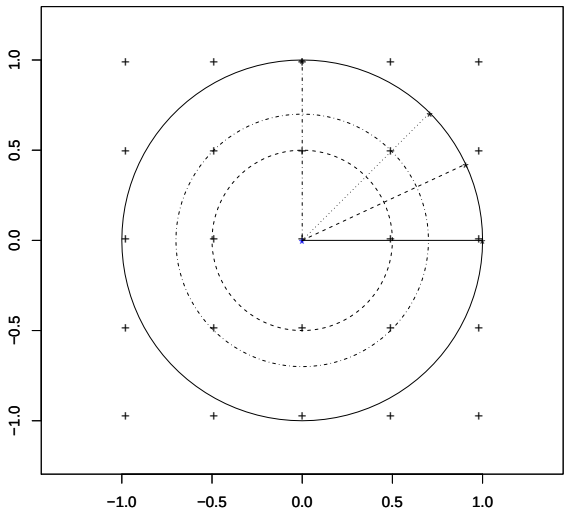

Figure 3: A plot of the two dimensional study region with the"+" sign denoting 25 knot-locations. Also shown are four angular directions at 0, 25, 45 and 60 degrees and three concentric circles with radii values 0.3, 0.5 and 1.0.

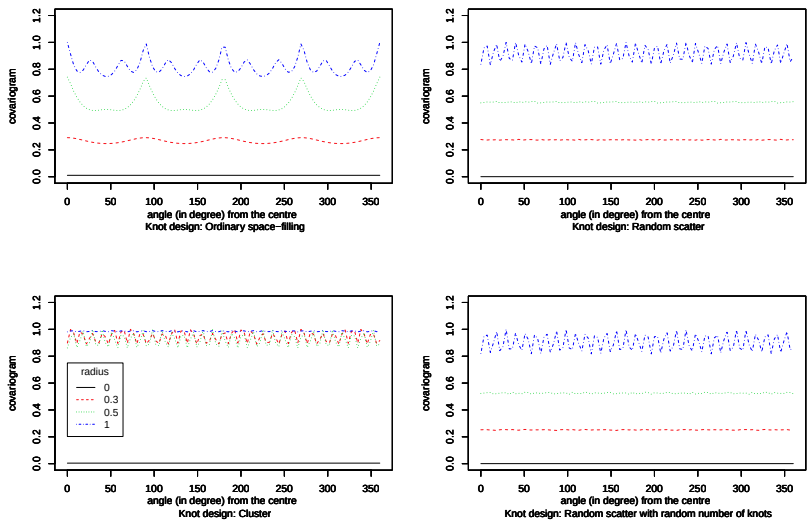

Figure 4: Semivariogram plots against angle for the two dimensional example under different knot configurations.

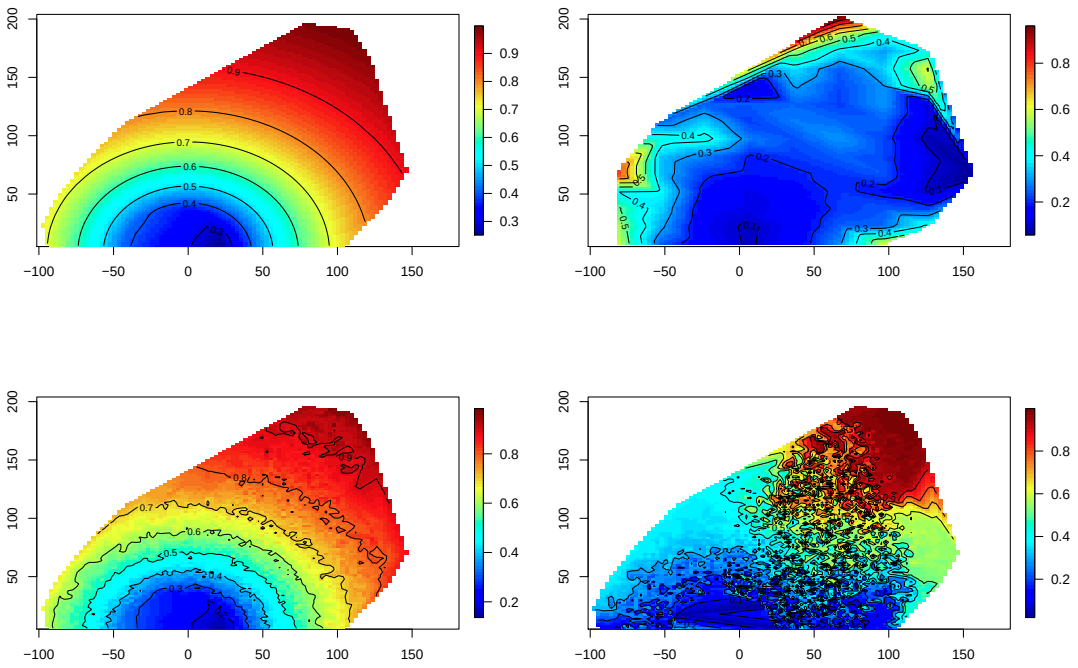

Figure 5: Semivariogram contour plots for the 1990 scallop data set. (a) Theoretical contours for a stationary and isotropic model. (b) ESC plot of the observed data. (c) TSC plot for a fixed space filling knot design with 100 knots. (d) TSC plot for a random space filling knot-design with 100 knots.

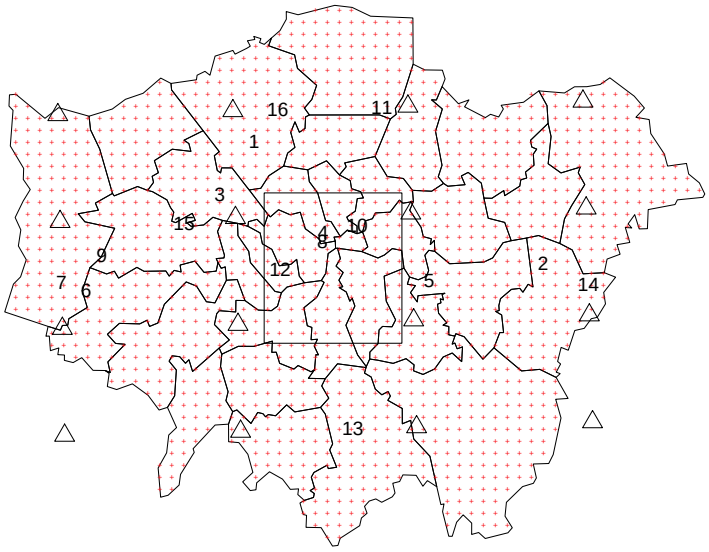

Figure 6: A map of greater London showing the monitoring sites as numbered texts, and the corners of the 1 kilometre square grid cells as red plus signs. Large triangles show the knot locations for the space filling design with 16 knots. The rectangular box in the middle represent the area where all the knots were assumed to be under the clustering design for knot selection.

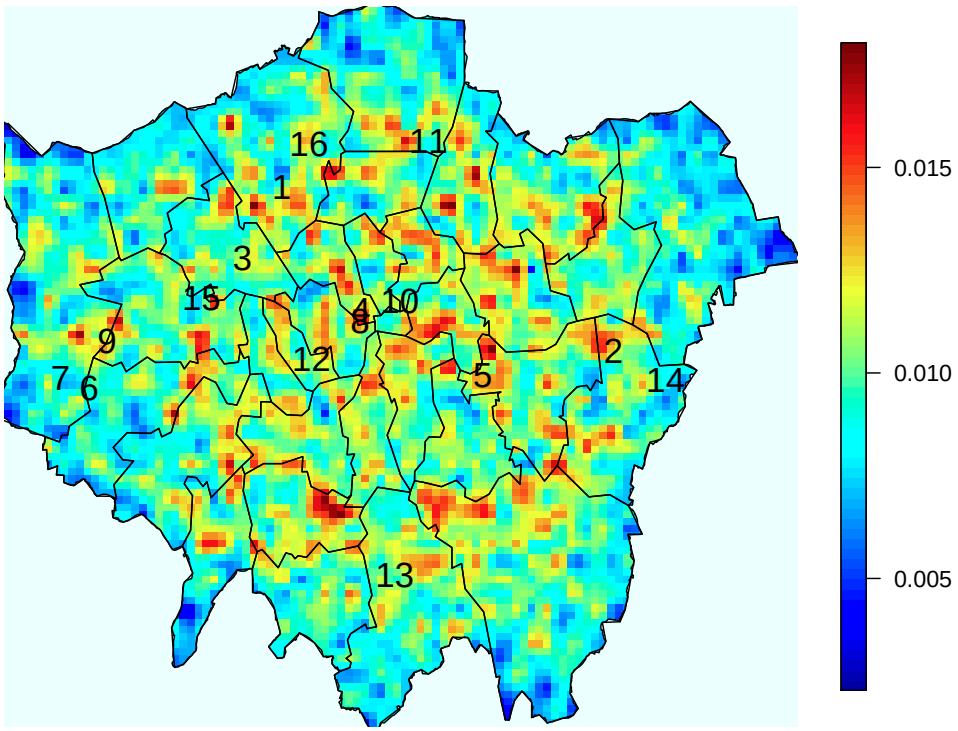

Figure 7: An interpolated surface showing the posterior probability of the knot-locations. The observation sites are superimposed.

		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	0.91	3.39	3.12	2.71	2.30	1.54
P	9.43	13.04	6.84	5.54	6.08	6.62
G+P	10.33	16.43	9.96	8.25	8.38	8.16
RMSPE	8.87	11.76	11.66	5.28	5.29	4.80
MAPE	6.74	10.24	8.35	4.39	4.42	3.93

Table 4: PMCC values and summaries of leave one out cross-validation values for models fitted to annual PM_{10} data from 16 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.

Parameter	GEM16	RSF25
β_0	–14.35 (–33.21, 14.51)	–20.03 (–29.93, –10.52)
β_1	5.72 (-5.76, 9.62)	7.39 (4.60, 10.31)
σ_w^2	0.35 (0.13, 0.75)	0.73 (0.51, 1.32)
σ_ϵ^2	0.35 (0.15, 0.82)	0.28 (0.20, 0.33)
ϕ	2.46 (0.06, 4.59)	1.51 (0.49, 2.36)

Table 5: Parameter estimates (posterior means) for all the parameters under different models fitted to annual PM_{10}, data from 16 monitoring sites in London for the year 2011. GEM16 stands for the full dimensional spatial random effects model. RSF25 denotes random space filling with 25 knots.

site	observed	predicted	sd	95% prediction limits
1	20.9	25.8	6.06	(15.89, 39.69)
2	21.0	26.1	6.09	(15.51, 37.18)
3	39.0	26.2	5.49	(17.36, 35.50)
4	22.6	26.3	6.62	(18.73, 42.53)
5	23.2	27.6	7.13	(15.93, 39.80)
6	17.8	24.5	5.96	(14.89, 35.37)
7	19.4	24.2	5.48	(13.81, 34.99)
8	31.5	27.9	6.65	(16.92, 41.25)
9	22.4	24.6	5.54	(14.10, 35.56)
10	31.4	28.7	6.21	(16.98, 41.81)
11	26.2	25.4	5.37	(15.04, 38.25)
12	26.8	28.0	6.01	(16.70, 40.64)
13	23.8	20.9	5.52	(11.17, 32.27)
14	26.3	24.3	5.46	(14.52, 35.75)
15	32.1	30.3	6.35	(15.94, 38.29)
16	27.7	24.4	6.54	(14.13, 36.58)

Table 6: Observed PM_{10} and leave one out cross-validation predictions at all the 16 data sites using the RSF25 model. Also shown are standard deviations and 95% prediction limits.