Non-Gaussian and Nonparametric Models for Continuous Spatial Data

Mark F.J. Steel and Montserrat Fuentes

December 8, 2008

Contents

Chapter 1

Non-Gaussian and Nonparametric Models for Continuous Spatial Data

Statistical modelling of continuous spatial data is often based on Gaussian processes. This typically facilitates prediction, but Normality is not necessarily an adequate modelling assumption for the data at hand. This has lead some authors to propose data transformations before using a Gaussian model: in particular, [1] propose to use the Box-Cox family of power transformations. An approach based on generalized linear models for spatial data is presented in [2]. In this chapter we present some flexible ways of modelling that allow the data to inform us on an appropriate distributional assumption. There are two broad classes of approaches we consider: firstly, we present a purely parametric modelling framework, which is wider than the Gaussian family, with the latter being a limiting case. This is achieved by scale mixing a Gaussian process with another process, and is particularly aimed at accommodating heavy tails. In fact, this approach allows us to identify spatial heteroscedasticity, and leads to relatively simple inference and prediction procedures. A second class of models is based on Bayesian nonparametric procedures. Most of the approaches discussed fall within the family of stick-breaking priors, which we will discuss briefly. These models are very flexible, in that they do not assume a single parametric family, but allow for highly non-Gaussian behaviour. A perhaps even more important property of the models

discussed in this chapter is that they accommodate nonstationary behaviour.

We will adopt a Bayesian framework throughout this chapter. In order to focus on the non-Gaussian properties in space, we shall only consider spatial processes. Extensions to spatio-temporal settings are, in principle, straightforward. Throughout, we denote a k-variate Normal distribution on a random vector $\mathbf{y}$ with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ as $\mathbf{y} \sim \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with density function $f_N^k(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1.1 Non-Gaussian Parametric Modelling

There are a number of parametric approaches leading to non-Gaussian models for continuous data that will not be discussed in detail in this section. In particular, we shall not deal with models that use a transformation of the data (as in [1]) to induce Gaussian behaviour. Another approach which is not discussed here is the use of Gaussian processes as a component within a nonlinear model for the observations, such as a generalized linear model (see [2] and further discussed in Chapter 2.3, Section 8) or a frailty model for survival data, such as used in [3] and [4]. In addition, we shall omit discussion of some application-specific ways of modelling non-Gaussian data, such as the approach of [5] to model the opacity of flocculated paper.

Let $Y(\mathbf{s})$ be a random process defined for locations $\mathbf{s}$ in some spatial region $\mathcal{D} \subset \Re^d$. We assume the model

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T\boldsymbol{\beta} + \eta(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{1.1}$$

where the mean surface is assumed to be a linear function of $\mathbf{x}(\mathbf{s})^T = (x_1(\mathbf{s}), \ldots, x_k(\mathbf{s}))$, a vector of k variables, which typically include known functions of the spatial coordinates, with unknown coefficient vector $\boldsymbol{\beta} \in \Re^k$. Further, $\eta(\mathbf{s})$ is a second-order stationary error process with zero mean and variance σ^2 and with an isotropic correlation function (depending only on the distance between points) $\mathrm{corr}\,[\eta(\mathbf{s}_i), \eta(\mathbf{s}_j)] = C_{\boldsymbol{\theta}}(\|\mathbf{s}_i - \mathbf{s}_j\|)$, where $C_{\boldsymbol{\theta}}(d)$ is a valid correlation function of distance d, parameterized by a vector $\boldsymbol{\theta}$. Finally, $\epsilon(\mathbf{s})$ denotes an uncorrelated Gaussian process with mean zero and variance τ^2, modelling the so-called

"nugget effect" (or "pure error", allowing for measurement error and small scale variation). The ratio $\omega^2 = \tau^2/\sigma^2$ indicates the relative importance of the nugget effect.

We assume that we have observed a single realization from this random process at n different locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and we denote the vector observation by $\mathbf{y} = (y_1, \ldots, y_n)^T$, where we use the notation $y_i = Y(\mathbf{s}_i)$. As mentioned above, the most commonly made distributional assumption is that $\eta(\mathbf{s})$ is a Gaussian process, which implies that $\mathbf{y}$ follows an n-variate Normal distribution with $E[\mathbf{y}] = X^T\boldsymbol{\beta}$, where $X = (\mathbf{x}(\mathbf{s}_1), \ldots, \mathbf{x}(\mathbf{s}_n))$, and $\text{var}[\mathbf{y}] = \sigma^2\mathbf{C}_{\boldsymbol{\theta}} + \tau^2\mathbf{I}_n$, where $\mathbf{C}_{\boldsymbol{\theta}}$ is the $n \times n$ correlation matrix with $C_{\boldsymbol{\theta}}(\|\mathbf{s}_i - \mathbf{s}_j\|)$ as its (i,j)th element. Note that even though we only have one observation per location, we are still able to criticize the normality assumption: in particular, the n elements of $B(\mathbf{y} - X^T\boldsymbol{\beta})$ where $B^{-1}(B^{-1})^T = \sigma^2\mathbf{C}_{\boldsymbol{\theta}} + \tau^2\mathbf{I}_n$ are assumed to be independent draws from a standard Normal given all model parameters.

1.1.1 The Gaussian-log-Gaussian mixture model

In [6] an alternative stochastic specification based on scale mixing the Gaussian process $\eta(\mathbf{s})$ is proposed. In particular, a mixing variable $\lambda_i \in \Re_+$ is assigned to each observation $i = 1, \ldots, n$, and the sampling model for the ith location, $i = 1, \ldots, n$, is now changed to

$$y_i = \mathbf{x}(\mathbf{s}_i)^T\boldsymbol{\beta} + \frac{\eta_i}{\sqrt{\lambda_i}} + \epsilon_i, \tag{1.2}$$

where we have used the notation $\eta_i = \eta(\mathbf{s}_i)$ and $\epsilon_i = \epsilon(\mathbf{s}_i)$, and $\epsilon_i \sim \text{N}_1(0, \tau^2)$, i.i.d. and independent of $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)' \sim \text{N}_n(\mathbf{0}, \sigma^2\mathbf{C}_{\boldsymbol{\theta}})$. The mixing variables λ_i are independent of ϵ_i and $\boldsymbol{\eta}$. In order to verify that the sampling model described above is consistent with a well-defined stochastic process, we can check that the Kolmogorov consistency conditions are satisfied. In [6] it is shown that (1.2) does support a stochastic process provided that the distribution of the mixing variables satisfies a very weak symmetry condition under permutation.

In this spatial model, scale mixing introduces a potential problem with the continuity of the resulting random field Y. Let us, therefore, consider a stationary process $\lambda(\mathbf{s})$ for

the mixing variables, so that $\lambda_i = \lambda(\mathbf{s}_i)$. The representation in (1.2) makes clear that we are now replacing the Gaussian stochastic process $\eta(\mathbf{s})$ by a ratio of independent stochastic processes $\eta(\mathbf{s})/\sqrt{\lambda(\mathbf{s})}$. Mean square continuity of the spatial process $\eta(\mathbf{s})/\sqrt{\lambda(\mathbf{s})}$ is defined by $\mathrm{E}[\{\eta(\mathbf{s}_i)/\sqrt{\lambda(\mathbf{s}_i)} - \eta(\mathbf{s}_j)/\sqrt{\lambda(\mathbf{s}_j)}\}^2]$ tending to zero as $\mathbf{s}_i \to \mathbf{s}_j$. Assuming that $\mathrm{E}[\lambda_i^{-1}]$ exists, we obtain

$$\mathrm{E}\left[\left\{\frac{\eta_i}{\sqrt{\lambda_i}} - \frac{\eta_j}{\sqrt{\lambda_j}}\right\}^2\right] = 2\sigma^2\left\{\mathrm{E}[\lambda_i^{-1}] - C_{\boldsymbol{\theta}}(\|\mathbf{s}_i - \mathbf{s}_j\|)\mathrm{E}[\lambda_i^{-1/2}\lambda_j^{-1/2}]\right\},$$

which in the limit as $\|\mathbf{s}_i - \mathbf{s}_j\| \to 0$ tends to $2\sigma^2\left\{\mathrm{E}[\lambda_i^{-1}] - \lim_{\|\mathbf{s}_i-\mathbf{s}_j\|\to 0}\mathrm{E}[\lambda_i^{-1/2}\lambda_j^{-1/2}]\right\}$. If λ_i and λ_j are independent, then $\lim_{\|\mathbf{s}_i-\mathbf{s}_j\|\to 0}\mathrm{E}[\lambda_i^{-1/2}\lambda_j^{-1/2}] = \{\mathrm{E}[\lambda_i^{-1/2}]\}^2 \leq \mathrm{E}[\lambda^{-1}]$ from Jensen's inequality and, thus, $\eta(\mathbf{s})/\sqrt{\lambda(\mathbf{s})}$ is not mean square continuous. This can also be seen immediately by considering the logarithm of the process $\log\{\eta(\mathbf{s})/\sqrt{\lambda(\mathbf{s})}\} = \log\{\eta(\mathbf{s})\} - (1/2)\log\{\lambda(\mathbf{s})\}$. This discontinuity essentially arises from the fact that two separate locations, no matter how close, are assigned independent mixing variables. Thus, in order to induce mean square continuity of the process (in the version without the nugget effect), we need to correlate the mixing variables in $\boldsymbol{\lambda}$, so that locations that are close will have very similar values of λ_i. In particular, if $\lambda^{-1/2}(\mathbf{s})$ is itself mean square continuous, then $\eta(\mathbf{s})/\sqrt{\lambda(\mathbf{s})}$ is a mean square continuous process.

Therefore, [6] consider the following mixing distribution:

$$\ln(\boldsymbol{\lambda}) = (\ln(\lambda_1), \ldots, \ln(\lambda_n))^T \sim \mathrm{N}_n\left(-\frac{\nu}{2}\mathbf{1}, \nu\mathbf{C}_{\boldsymbol{\theta}}\right), \tag{1.3}$$

where $\mathbf{1}$ is a vector of ones, and we correlate the elements of $\ln(\boldsymbol{\lambda})$ through the same correlation matrix as $\boldsymbol{\eta}$. Equivalently, we assume a Gaussian process for $\ln(\lambda(\mathbf{s}))$ with constant mean surface at $-\nu/2$ and covariance function $\nu\mathbf{C}_{\boldsymbol{\theta}}(\|\mathbf{s}_i - \mathbf{s}_j\|)$. One scalar parameter $\nu \in \Re_+$ is introduced in (1.3), which implies a log-Normal distribution for λ_i with $\mathrm{E}[\lambda_i] = 1$ and $\mathrm{var}[\lambda_i] = \exp(\nu) - 1$. Thus, the marginal distribution of λ_i is tight around unity for very small ν (of the order $\nu = 0.01$) and as ν increases, the distribution becomes more spread out and more right skewed, while the mode shifts towards zero. For example, for $\nu = 3$, the variance is 19.1 and there is a lot of mass close to zero. It is exactly values of λ_i close

to zero that will lead to an inflation of the scale in (1.2) and will allow us to accommodate heavy tails. On the other hand, as $\nu \to 0$, we retrieve the Gaussian model as a limiting case. Figure 1.1 illustrates this behaviour.

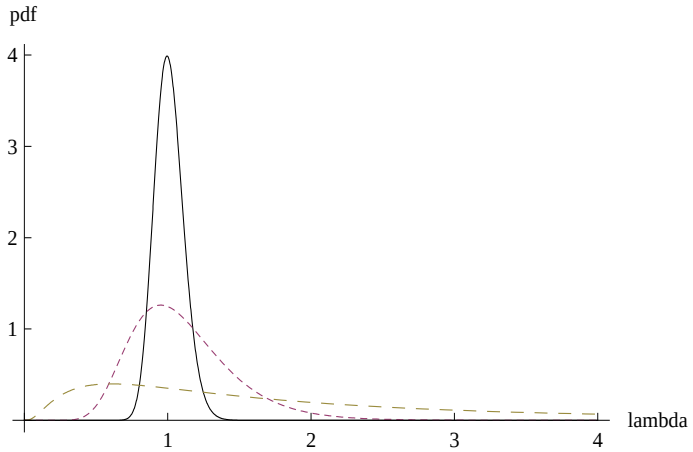


Figure 1.1: Marginal probability density function of mixing variables λ_i for various values of ν. Solid line $\nu = 0.01$; short dashes: $\nu = 0.1$; long dashes: $\nu = 1$.

In [6] the mixture model defined by (1.2) and (1.3) is called the Gaussian-log-Gaussian (GLG) model. This approach is similar to that of [7] and [8], where an additional space deformation as in [9] is used to introduce non-stationarity. Note that the latter complication requires repeated observations for reliable inference.

1.1.2 Properties and interpretation

The correlation structure induced by the GLG model is given by

$$\text{corr}\,[y_i, y_j] = C_{\boldsymbol{\theta}}(\|\mathbf{s}_i - \mathbf{s}_j\|)\frac{\exp\left(\nu\left\{1 + \frac{1}{4}[C_{\boldsymbol{\theta}}(\|\mathbf{s}_i - \mathbf{s}_j\|) - 1]\right\}\right)}{\exp(\nu) + \omega^2}. \tag{1.4}$$

Thus, in the case without nugget effect ($\omega^2 = 0$) we see that if the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$ tends to zero, the correlation between y_i and y_j tends to one, so that the mixing does not induce a discontinuity at zero. It can also be shown (see [6]) that the smoothness of the process is not affected by the mixing, in the sense that without the nugget effect the process $Y(\mathbf{s})$ has exactly the same smoothness properties as $\eta(\mathbf{s})$.

The tail behaviour of the finite-dimensional distributions induced by the GLG process is determined by the extra parameter ν. In particular, [6] derive that the kurtosis of the marginal distributions is given by $\text{kurt}[y_i] = 3\exp(\nu)$, again indicating that large ν corresponds to heavy tails, and Gaussian tails are the limiting case as $\nu \to 0$.

Our chosen specification for mixing the spatially dependent process as in (1.2) requires a smooth $\lambda(\mathbf{s})$ process which means that observations with particularly small values of λ_i will tend to cluster together. Thus, what we are identifying through small values of λ_i are regions of the space where the observations tend to be relatively far away from the estimated mean surface. Therefore, we can interpret the presence of relatively small values of λ_i in terms of spatial heteroskedasticity, rather than the usual concept of outlying observations. However, for convenience we will continue to call observations with small λ_i "outliers".

It may be useful to have an indication of which areas of the space require an inflated variance. Indicating regions of the space where the Gaussian model fails to fit the data well might suggest extensions to the underlying trend surface (such as missing covariates) that could make a Gaussian model a better option. The distribution of λ_i is informative about the outlying nature of observation i. Thus, [6] propose to compute the ratio between the posterior and the prior density functions for λ_i evaluated at $\lambda_i = 1$, *i.e.*

$$R_i = \frac{p(\lambda_i|\mathbf{y})}{p(\lambda_i)}|_{\lambda_i=1}\,. \tag{1.5}$$

In fact, this ratio R_i is the so-called Savage-Dickey density ratio, which would be the Bayes factor in favour of the model with $\lambda_i = 1$ (and all other elements of $\boldsymbol{\lambda}$ free) versus the model with free λ_i (*i.e.* the full mixture model proposed here) if $C_{\boldsymbol{\theta}}(\|\mathbf{s}_i - \mathbf{s}_j\|) = 0$ for all $j \neq i$. In this case, the Savage-Dickey density ratio is not the exact Bayes factor, but has to be adjusted as in [10]. The precise adjustment in this case is explained in [6]. Bayes factors convey the relative support of the data for one model versus another and immediately translate into posterior probabilities of rival models since the posterior odds (the ratio of two posterior model probabilities) equals the Bayes factor times the prior odds (the ratio of the prior model probabilities).

1.1.3 Prediction

An important reason for geostatistical modelling is prediction at unsampled sites. We wish to fully incorporate all uncertainty in the problem, including the covariance function.

Let $\mathbf{y} = (\mathbf{y}_o^T, \mathbf{y}_p^T)^T$ where $\mathbf{y}_o$ correspond to the $n-f$ observed locations and $\mathbf{y}_p$ is a vector of values to predict at f given sites. We are interested in the posterior predictive distribution of $\mathbf{y}_p$, *i.e.*

$$p(\mathbf{y}_p|\mathbf{y}_o) = \int p(\mathbf{y}_p|\mathbf{y}_o, \boldsymbol{\lambda}, \boldsymbol{\zeta}) p(\boldsymbol{\lambda}_p|\boldsymbol{\lambda}_o, \boldsymbol{\zeta}, \mathbf{y}_o) p(\boldsymbol{\lambda}_o, \boldsymbol{\zeta}|\mathbf{y}_o) \mathrm{d}\boldsymbol{\lambda} \mathrm{d}\boldsymbol{\zeta}, \tag{1.6}$$

where we have partitioned $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_o^T, \boldsymbol{\lambda}_p^T)^T$ conformably with $\mathbf{y}$ and $\boldsymbol{\zeta} = (\boldsymbol{\beta}, \sigma, \tau, \boldsymbol{\theta}, \nu)$. The integral in (1.6) will be approximated by Monte Carlo simulation, where the draws for $(\boldsymbol{\lambda}_o, \boldsymbol{\zeta})$ are obtained directly from the Markov chain Monte Carlo (MCMC) inference algorithm (which is described in some detail in [6]), and, since $p(\boldsymbol{\lambda}_p|\boldsymbol{\lambda}_o, \boldsymbol{\zeta}, \mathbf{y}_o) = p(\boldsymbol{\lambda}_p|\boldsymbol{\lambda}_o, \nu)$ we can evaluate (1.6) by using drawings for $\boldsymbol{\lambda}_p$ from

$$p(\ln \boldsymbol{\lambda}_p|\boldsymbol{\lambda}_o, \nu) = f_N^f \left(\ln \boldsymbol{\lambda}_p \,|\, \frac{\nu}{2}[\mathbf{C}_{po}\mathbf{C}_{oo}^{-1}\mathbf{1}_n - \mathbf{1}_f] + \mathbf{C}_{po}\mathbf{C}_{oo}^{-1} \ln \boldsymbol{\lambda}_o,\, \nu[\mathbf{C}_{pp} - \mathbf{C}_{po}\mathbf{C}_{oo}^{-1}\mathbf{C}_{op}] \right), \tag{1.7}$$

where we have partitioned

$$\mathbf{C}_{\boldsymbol{\theta}} = \begin{pmatrix} \mathbf{C}_{oo} & \mathbf{C}_{op} \\ \mathbf{C}_{po} & \mathbf{C}_{pp} \end{pmatrix}$$

conformably with $\mathbf{y}$. Thus, for each posterior drawing of $(\boldsymbol{\lambda}_o, \boldsymbol{\zeta})$, we will generate a drawing from (1.7) and evaluate

$$p(\mathbf{y}_p|\mathbf{y}_o, \boldsymbol{\lambda}, \boldsymbol{\zeta}) = f_N^f \left(\mathbf{y}_p | (\mathbf{X}_p - \mathbf{A}\mathbf{X}_o)\boldsymbol{\beta} + \mathbf{A}\mathbf{y}_o, \sigma^2 \left(\boldsymbol{\Lambda}_p^{-\frac{1}{2}} \mathbf{C}_{pp} \boldsymbol{\Lambda}_p^{-\frac{1}{2}} + \omega^2 \mathbf{I}_f - \mathbf{A}\boldsymbol{\Lambda}_o^{-\frac{1}{2}} \mathbf{C}_{op} \boldsymbol{\Lambda}_p^{-\frac{1}{2}} \right) \right), \tag{1.8}$$

where $\mathbf{I}_f$ is the f-dimensional identity matrix, $\mathbf{A} = \boldsymbol{\Lambda}_p^{-\frac{1}{2}} \mathbf{C}_{po} \boldsymbol{\Lambda}_o^{-\frac{1}{2}} \left[\boldsymbol{\Lambda}_o^{-\frac{1}{2}} \mathbf{C}_{oo} \boldsymbol{\Lambda}_o^{-\frac{1}{2}} + \omega^2 \mathbf{I}_n \right]^{-1}$ and $\mathbf{X}$ and $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$ are partitioned conformably to $\mathbf{y}$. Averaging the densities in (1.8) will give us the required posterior predictive density function.

1.1.4 Correlation function and prior distribution

For the correlation function $C_{\boldsymbol{\theta}}(d)$, where d is the Euclidean distance, we use the flexible Matérn class:

$$C_{\boldsymbol{\theta}}(d) = \frac{1}{2^{\theta_2-1}\Gamma(\theta_2)} \left(\frac{d}{\theta_1}\right)^{\theta_2} \mathcal{K}_{\theta_2}\left(\frac{d}{\theta_1}\right), \tag{1.9}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ with $\theta_1 > 0$ the range parameter and $\theta_2 > 0$ the smoothness parameter and where $\mathcal{K}_{\theta_2}(\cdot)$ is the modified Bessel function of the third kind of order θ_2. As a consequence $\eta(\mathbf{s})$ and thus $Y(\mathbf{s})$ are q times mean square differentiable if and only if $\theta_2 > q$.

In order to complete the Bayesian model, we now need to specify a prior distribution for the parameters $(\boldsymbol{\beta}, \sigma^{-2}, \omega^2, \nu, \boldsymbol{\theta})$. The prior distribution used in [6] is a carefully elicited proper prior. They argue against the use of reference priors as used in [11] for a simpler Gaussian model with fixed smoothness parameter θ_2 and without the nugget effect. In addition, such a reference prior would be extremely hard to derive for the more general GLG model discussed here. The prior used has a product structure with a Normal prior for $\boldsymbol{\beta}$, a Gamma prior for σ^{-2}, and a generalised inverse Gaussian (GIG) prior for both ω^2 and ν. The prior on $\boldsymbol{\theta}$ either imposes prior independence between θ_1 and θ_2 or between the alternative range parameter $\rho = 2\theta_1\sqrt{\theta_2}$ (see [12], p.51) and θ_2. In both cases, the prior on the Matérn parameters consists of the product of two judiciously chosen Exponential distributions.

An extensive sensitivity analysis in [6] suggests that a data set of small (but typical) size is not that informative on certain parameters. In particular, the parameters $(\nu, \boldsymbol{\theta}, \omega^2)$ are not that easily determined by the data and thus require very careful prior elicitation. In general, spatial models do suffer from weak identification issues and, thus, prior specification is critical. Chapter 2.3 (Section 4) discusses the fact that some parameters are not consistently estimated by classical maximum likelihood methods under infill asymptotics.

1.1.5 An application to Spanish temperature data

We analyse the maximum temperatures recorded in an unusually hot week in May 2001 in 63 locations within the Spanish Basque country. So $\mathcal{D} \subset \Re^2$ and for the trend function $\mathbf{x}(\mathbf{s})$

we use a quadratic form in the coordinates (with linear terms and the cross-product). As this region is quite mountainous (with the altitude of the monitoring stations in between 16 and 1188 meters), altitude is added as an extra explanatory variable (corresponding to regression coefficient β_7). Table 1.1 presents some posterior results for $\boldsymbol{\beta}$, using both the Gaussian and the GLG model. The Gaussian model tends to higher absolute values for β_2 and

Model	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Gaussian	3.19(0.22)	-0.20(0.23)	0.19(0.31)	-0.20(0.23)	0.37(0.45)	-0.24(0.28)	-0.40(0.18)
GLG	3.23(0.06)	-0.08(0.11)	0.12(0.13)	-0.19(0.09)	0.09(0.24)	-0.17(0.14)	-0.42(0.07)

Table 1.1: Temperature data: Posterior means (standard deviation) of the trend parameters.

β_5 and the inference on $\boldsymbol{\beta}$ is generally a lot less concentrated for this model. In both models higher altitude tends to reduce the mean temperature, as expected. Posterior inference

	Gaussian	GLG
σ	0.32 (0.11)	0.09 (0.03)
ω^2	1.22 (0.79)	1.27 (1.12)
τ	0.31 (0.06)	0.08 (0.02)
θ_1	5.71 (10.33)	4.02 (12.70)
θ_2	1.87 (2.03)	0.61 (0.98)
ρ	8.64 (8.20)	2.35 (2.97)
ν	0 (0)	2.51 (0.76)
$\sigma^2 \exp(\nu)$	0.11 (0.09)	0.12 (0.15)
$\tau/[\sigma \exp(\nu/2)]$	1.05 (0.35)	0.30 (0.12)

Table 1.2: Temperature data: Posterior means (standard deviation) for some non-trend parameters.

on the other parameters in the models is presented in Table 1.2. Clearly, the Gaussian model assigns a larger importance to the nugget effect (see the difference in $\tau/[\sigma \exp(\nu/2)]$, which is the ratio of standard deviations between the process inducing the nugget effect and the spatial process), while making the surface a lot smoother than the GLG model. In order to accommodate the outlying observations (see later), the Gaussian model needs to dramatically increase the values of both σ and τ. Since most of the posterior mass for ν is well away from zero, it is not surprising that the evidence in favour of the GLG model is very strong indeed. In particular, the Bayes factor in favour of the GLG model is $3.4 \cdot 10^{20}$, a lot of which is attributable to three very extreme observations: observations 20, 36 and 40, which are all close together. Table 1.3 presents the Bayes factors in favour of $\lambda_i = 1$ for the

four observations with smallest mean λ_i. The column labelled "corr" is the multiplicative correction factor to the Savage-Dickey density ratio mentioned in Subsection 1.1.2. Clearly,

obs.#	$E[\lambda_i\|\mathbf{z}]$	S.Dev.$[\lambda_i\|\mathbf{z}]$	(1.5)	corr	BF for $\lambda_i = 1$
20	0.020	0.024	0.000	0.74	0.000
36	0.015	0.020	0.000	0.79	0.000
40	0.016	0.020	0.000	0.62	0.000
41	0.059	0.085	0.006	0.57	0.004

Table 1.3: Temperature data: Bayes factors in favour of $\lambda_i = 1$ for selected observations. The entries 0.000 indicate values less than 0.0005.

all observations listed in Table 1.3 are outliers, indicating two regions with inflated variance.

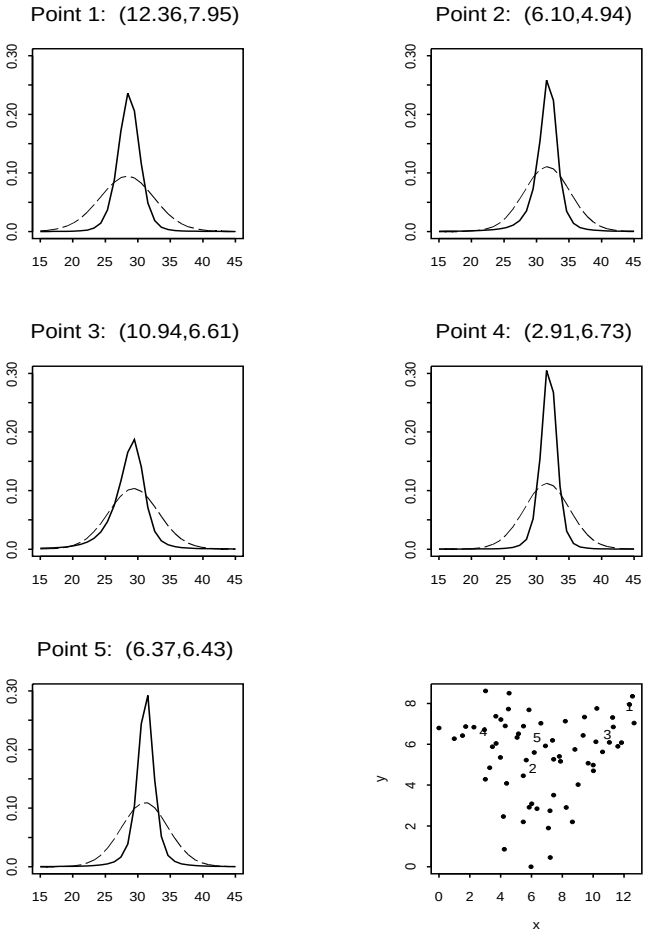


Figure 1.2: Temperature data: Predictive densities at five unobserved locations. The observables are measured in degrees centigrade, and the elevations at the predicted sites range from 53 (point 2) to 556 (point 3) meters. Dashed line: Gaussian; solid line: GLG. The lower right panel indicates the locations of the observed sites by dots and the five unobserved sites by their respective numbers.

Figure 1.2 displays the predictive densities, computed as in Subsection 1.1.3 for five unobserved locations, ranging in altitude from 53 to 556 meters. The GLG model leads to heavier extreme tails than the Gaussian model as a consequence of the scale mixing. Nevertheless, in the (relevant) central mass of the distribution, the GLG predictives clearly are more concentrated than the Gaussian ones, illustrating that the added uncertainty due to the scale mixing is more than offset by changes in the inference on other aspects of the model. In particular, the nugget effect is much less important for the non-Gaussian model. From (1.8) it is clear that the predictive standard deviation is bounded from below by τ (in order to interpret the numbers in Table 1.2 in terms of observables measured in degrees centigrade, we need to multiply τ by a factor 10, due to scaling of the data). Clearly, a lot of the predictive uncertainty in the Gaussian case is due to the nugget effect.

1.2 Bayesian Nonparametric Approaches

In the next section we will use nonparametric models for the spatial components, which can accommodate much more flexible forms and can also easily deal with skewness, multimodality etc. In addition, even though the prior predictive distributions induced by these models are stationary, the posterior predictives can accommodate very nonstationary behaviour. As the Bayesian nonparametric methods presented are all based on the broad class of stick-breaking priors, we will first briefly explain this class of priors. See [13] for an excellent overview of nonparametric Bayesian inference procedures, while [14] provides an insightful and very up-to-date discussion of nonparametric Bayesian methods, specifically aimed at applications in biostatistics. One approach which we will not discuss here is that of transforming the space corresponding to a Gaussian parametric model, as introduced in [9] and developed in [15] in a Bayesian framework.

1.2.1 Stick-breaking priors

Bayesian nonparametric methods avoid dependence on parametric assumptions by working with probability models on function spaces, in other words, by using (in principle) infinitely

many parameters. A useful and broad class of such random probability measures is the class of stick-breaking priors. This class was discussed in some detail by [16] and is at the basis of many recent studies.

A random probability distribution, F, has a stick-breaking prior if

$$F \stackrel{d}{=} \sum_{i=1}^{N} p_i \delta_{\boldsymbol{\theta}_i}, \tag{1.10}$$

where δ_z denotes a Dirac measure at z, $p_i = V_i \prod_{j<i}(1 - V_j)$ where $V_1, \ldots, V_{N-1}$ are independent with $V_i \sim \text{Beta}(a_i, b_i)$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ are independent draws from a centring (or base) distribution H.

The definition in (1.10) allows for either finite or infinite N (with the latter corresponding to the conventional definition of nonparametrics). For $N = \infty$ several interesting and well-known processes fall into this class:

1. The Dirichlet process prior (see [17]) characterised by MH, where M is a positive scalar (often called the mass parameter), arises when V_i follows a $\text{Beta}(1, M)$ for all i. This representation was first given by [18].
2. The Pitman-Yor (or two-parameter Poisson-Dirichlet) process occurs if V_i follows a $\text{Beta}(1 - a, b + ai)$ with $0 \leq a < 1$ and $b > -a$. As special cases we can identify the Dirichlet process for $a = 0$ and the stable law when $b = 0$.

Stick-breaking priors such as the Dirichlet process almost surely lead to discrete probability distributions. This is often not desirable for directly modelling observables that are considered realizations of some continuous process. To avoid this problem, the mixture of Dirichlet process model (introduced in [19]) is now the most commonly used specification in practice. Such models assume a continuous model for the observables, given some unknown parameters, and then use a stick-breaking prior as in (1.10) to model these parameters nonparametrically.

An important aspect of these models is that they tend to cluster the observations by assigning several observations to the same parameter values (or atoms of the nonparametric

distribution).

Conducting inference with such models relies on MCMC computational methods. One approach corresponds to marginalising out F and using a Polya urn representation to conduct a Gibbs sampling scheme. See [20] for a detailed description of such methods. Another approach (see [16]) directly uses the stick-breaking representation in (1.10) and either truncates the sum or avoids truncation through slice sampling or the retrospective sampler proposed in [21]. An accessible and more detailed discussion of computational issues can be found in *e.g.* [14].

In order to make this wide class of nonparametric priors useful for our spatial context, we need to somehow index it by space. More generally, we can attempt to introduce dependencies on time or other covariates (leading to nonparametric regression models). Most of the (rather recent) literature in this area follows the ideas in [22], who considered allowing the masses, $\mathbf{V} = (V_1, V_2, \ldots)$, or the locations, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots)$, of the atoms to follow a stochastic process defined over the domain. This leads to so-called Dependent Dirichlet (DDP) processes and a lot of this work concentrates on the "single-p" DDP model where only the locations, $\boldsymbol{\theta}$, follow stochastic processes. An application to spatial modelling is developed in [23] by allowing the locations $\boldsymbol{\theta}$ to be drawn from a random field (a Gaussian process). A generalization of this idea is briefly explained in the next subsection.

1.2.2 Generalized Spatial Dirichlet process

The idea in [23] is to introduce a spatial dependence through the locations, by indexing $\boldsymbol{\theta}$ with the location $\mathbf{s}$ and making $\boldsymbol{\theta}(\mathbf{s})$ a realization of a random field, with H being a stationary Gaussian process. Continuity properties of these realizations will then follow from the choice of covariance function. In the simple model $Y(\mathbf{s}) = \eta(\mathbf{s}) + \epsilon(\mathbf{s})$ where $\eta(\mathbf{s})$ has this spatial Dirichlet prior and $\epsilon(\mathbf{s}) \sim \mathrm{N}(0, \tau^2)$ is a nugget effect, the joint density of the observables $\mathbf{y} = [Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)]^T$ is almost surely a location mixture of Normals with density function of the form $\sum_{i=1}^{N} p_i f_N^n(\mathbf{y}|\boldsymbol{\eta}_i, \tau^2\mathbf{I}_n)$, using (1.10). This allows for a large amount of flexibility and is used in [23] to analyze a data set of precipitation data consisting of 75 replications at

39 observed sites in the South of France.

However, the joint distribution over any set of locations uses the same set of weights $\{p_i\}$, so the choice between the random surfaces in the location mixture is not dependent on location. In [24] and [25] this framework is extended to allow for the surface selection to vary with the location, while still preserving the property that the marginal distribution at each location is generated from the usual Dirichlet process. This extension, called the generalized spatial Dirichlet process model, assumes that the random probability measure on the space of distribution functions for $\eta(\mathbf{s}_1), \ldots, \eta(\mathbf{s}_n)$ is

$$F^{(n)} \stackrel{d}{=} \sum_{i_1=1}^{\infty} \ldots \sum_{i_n=1}^{\infty} p_{i_1,\ldots,i_n} \delta_{\theta_{i_1}} \ldots \delta_{\theta_{i_n}}, \tag{1.11}$$

where $i_j = i(s_j), j = 1, \ldots, n$, the locations θ_{i_j} are drawn from the centering random field H and the weights $p_{i_1,\ldots,i_n}$ are distributed independently from the locations on the infinite unit simplex. These weights allow for the site-specific selection of surfaces and are constrained to be consistent (to define a proper random process) and continuous (in the sense that they assign similar weights for sites that are close together). This will induce a smooth (mean-square continuous) random probability measure.

¿From (1.11) it is clear that we can think of this generalization in terms of a multivariate stick-breaking representation. In [24] a particular specification for $p_{i_1,\ldots,i_n}$ is proposed, based on thresholding of auxiliary Gaussian random fields. This leads to processes that are non-Gaussian and nonstationary, with non-homogeneous variance.

As the locations associated with the sampled sites effectively constitute one single observation from the random field, we need replications in order to conduct inference using this approach. However, replications over time do not need to be independent and a dynamic model can be used.

1.2.3 Hybrid Dirichlet mixture models

The idea of the previous subsection is further developed in [26] in the context of functional data analysis. Here, we have observations of curves (*e.g.* in space or in time) and often it is important to represent the n observed curves by a smaller set of canonical curves. The curves $\mathbf{y}_i = [Y_i(\mathbf{x}_1), \ldots, Y_i(\mathbf{x}_m)]^T, i = 1, \ldots, n$ are assumed to be observed at a common set of coordinates $\mathbf{x}_1, \ldots, \mathbf{x}_m$.

[26] start from a more general class than stick-breaking priors, namely the species sampling prior which can be represented as (1.10) with a general specification on the weights. This leads them to a different way of selecting the multivariate weights in (1.11). In [26] these weights are interpreted as the distribution of a random vector of labels, assigning curves to locations. They use mixture modelling where the observations are normally distributed with a nonparametric distribution on the locations, *i.e.*

$$\begin{aligned} \mathbf{y}_i \mid \boldsymbol{\theta}_i &\overset{ind}{\sim} \mathrm{N}_m(\boldsymbol{\theta}_i, \sigma^2 \mathbf{I}_m), \\ \boldsymbol{\theta}_i \mid F_{\mathbf{x}_1,\ldots,\mathbf{x}_m} &\overset{i.i.d.}{\sim} F_{\mathbf{x}_1,\ldots,\mathbf{x}_m}, \end{aligned} \tag{1.12}$$

where

$$F_{\mathbf{x}_1,\ldots,\mathbf{x}_m} \overset{d}{=} \sum_{j_1=1}^{k} \cdots \sum_{j_m=1}^{k} p(j_1, \ldots, j_m)\, \delta_{\theta_{j_1,1},\ldots,\theta_{j_m,m}}, \tag{1.13}$$

where $p(j_1, \ldots, j_m)$ represents the proportion of (hybrid) species $(\theta_{j_1,1}, \ldots, \theta_{j_m,m})$ in the population, $p(j_1, \ldots, j_m) \geq 0$, $\sum_{j_1=1}^{k} \cdots \sum_{j_m=1}^{k} p(j_1, \ldots, j_m) = 1$, and $\boldsymbol{\theta}_j = (\theta_{j,1}, \ldots, \theta_{j,m}) \overset{i.i.d.}{\sim} H$ (again typically chosen to be an m-dimensional distribution of a Gaussian process), independently of the $p(j_1, \ldots, j_m)$'s.

This generates a location mixture of normals with local random effects. Hybrid DP mixtures are obtained as limits of the finite mixture framework above for $k \to \infty$. Functional dependence in the (hidden) label process is modelled through an auxiliary Gaussian copula, which contributes to the simplicity and the flexibility of the approach.

An application to MRI brain images in [26] illustrates the modeling of the species recombination (hybridization) through the labeling prior and the improvement over simple

mixtures of Dirichlet processes.

1.2.4 Order-Based Dependent Dirichlet process

An alternative approach to extending the framework in (1.10) is followed by [27], who define the ranking of the elements in the vectors $\mathbf{V}$ and $\boldsymbol{\theta}$ through an ordering $\boldsymbol{\pi}(\mathbf{s})$, which changes with the spatial index (or other covariates). Since weights associated with atoms that appear earlier in the stick-breaking representation tend to be larger (i.e. $\mathrm{E}[p_i(\mathbf{s})] < \mathrm{E}[p_{i-1}(\mathbf{s})]$), this induces similarities between distributions corresponding to similar orderings. The similarity between $\boldsymbol{\pi}(\mathbf{s}_1)$ and $\boldsymbol{\pi}(\mathbf{s}_2)$ will control the correlation between $F_{\mathbf{s}_1}$ and $F_{\mathbf{s}_2}$, the random distributions at these spatial locations. The induced class of models is called order-based dependent Dirichlet processes (πDDP).

This specification also preserves the usual Dirichlet process for the marginal distribution at each location, but, in contrast with the single-p approaches, leads to local updating, where the influence of observations decreases as they are further away.

The main challenge is to define stochastic processes $\boldsymbol{\pi}(\mathbf{s})$, and [27] use a point process Φ and a sequence of sets $U(\mathbf{s})$, which define the region in which points are relevant for determining the ordering at $\mathbf{s}$. The ordering, $\boldsymbol{\pi}(\mathbf{s})$, then satisfies the condition

$$\|\mathbf{s} - \mathbf{z}_{\pi_1(\mathbf{s})}\| < \|\mathbf{s} - \mathbf{z}_{\pi_2(\mathbf{s})}\| < \|\mathbf{s} - \mathbf{z}_{\pi_3(\mathbf{s})}\| < \dots,$$

where $\|\cdot\|$ is a distance measure and $\mathbf{z}_{\pi_i(\mathbf{s})} \in \Phi \cap U(\mathbf{s})$. We assume there are no ties, which is a.s. the case for *e.g.* Poisson point processes. Associating each atom $(V_i, \boldsymbol{\theta}_i)$ with the element of the point process $\mathbf{z}_i$ defines a marked point process from which we can define the distribution $F_{\mathbf{s}}$ for any $\mathbf{s} \in \mathcal{D}$. Using a stationary Poisson process for Φ, the autocorrelation function between random probability measures at different locations can be expressed in the form of deterministic integrals, as explained in [27]. More specific constructions can even lead to analytical expressions for the autocorrelation structure.

In particular, [27] define a practical proposal for spatial models through the so-called

permutation construction. This is obtained through defining $\mathcal{D} \subset \Re^d$ and $U(\mathbf{s}) = \mathcal{D}$ for all values of $\mathbf{s}$. In one dimension ($d = 1$), we can derive an analytic form for the autocorrelation function. Let Φ be Poisson with intensity λ, $\mathcal{D} \subset \Re$ and $U(s) = \mathcal{D}$ for all s. Then we obtain

$$\text{corr}(F_{s_1}, F_{s_2}) = \left(1 + \frac{2\lambda h}{M+2}\right) \exp\left\{\frac{-2\lambda h}{M+1}\right\},$$

where $h = |s_1 - s_2|$ is the distance between s_1 and s_2, and M is the mass parameter of the marginal Dirichlet process.

Note the unusual form of the correlation structure above. It is the weighted sum of a Matérn correlation function with smoothness parameter 3/2 (with weight $(M+1)/(M+2)$) and an exponential correlation function (with weight $1/(M+2)$), which is a less smooth member of the Matérn class, with smoothness parameter 1/2. So for $M \to 0$ the correlation function will tend to the arithmetic average of both and for large M the correlation structure will behave like a Matérn with smoothness parameter 3/2. In higher dimensions, for $d \geq 2$, the autocorrelation function can be expressed as a two-dimensional integral, as detailed in [27].

[27] suggest prior distributions for (M, λ) and use the order-based dependent Dirichlet process with the permutation construction to analyse the Spanish temperature data as described in Subsection 1.1.5. In particular, the model used is (1.1) with the sum of $\eta(\mathbf{s})$ and the intercept (β_1) modelled through a πDDP process, using as centring distribution H a $\text{N}(\bar{y}, \tau^2/\kappa)$, where $\bar{y}$ is the observation mean and an inverted Gamma prior is adopted for κ. In Figure 1.3 we display the posterior predictive distributions at the same unsampled locations as in Figure 1.2. The lower right panel indicates the location of these unobserved locations (with numbers), as well as the observed ones (with dots). While some of the predictives are similar to those obtained with the parametric GLG model, there is now clearly a much larger variety of predictive shapes, with multimodality and skewness illustrating the large flexibility of such nonparametric approaches. Of course, the πDDP part of the model does not only allow for departures from Gaussianity, but also serves to introduce the spatial correlation.

Finally, note that we do not require replication of observations at each site for inference

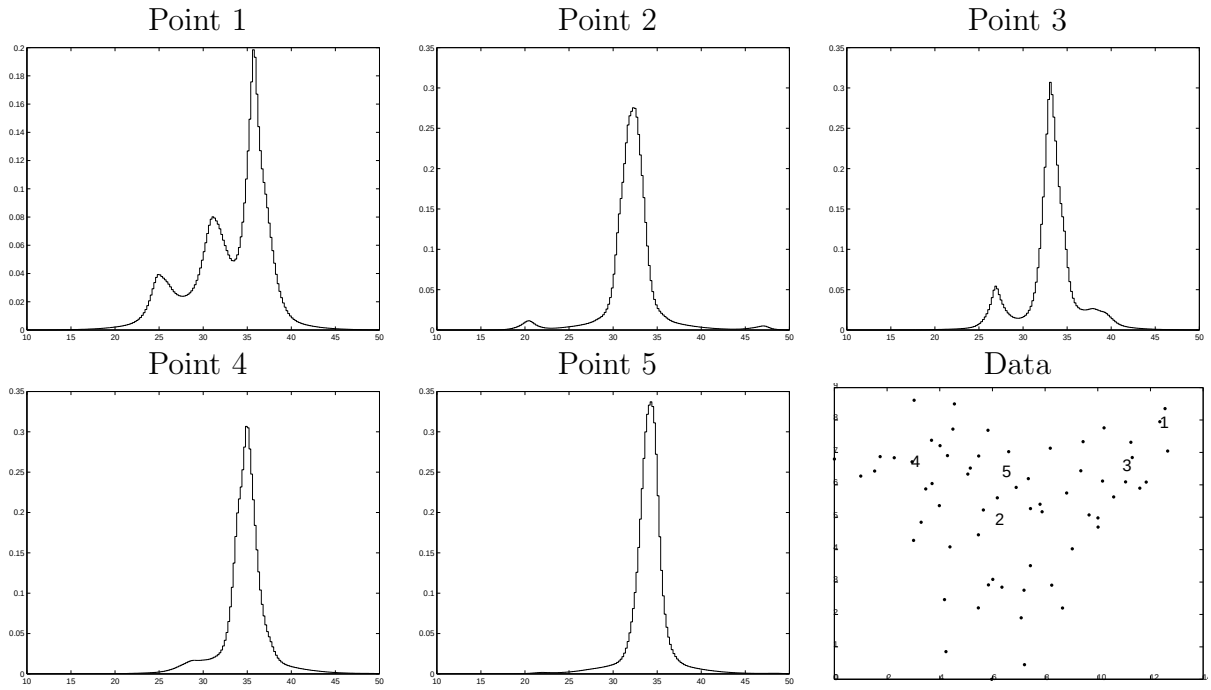


Figure 1.3: Temperature data: The posterior predictive distribution at five unobserved locations. The latter are indicated by numbers in the lower right-hand panel, where the observed locations are denoted by dots.

with πDDP models. In principle, extensions to spatiotemporal models can be formulated easily by treating time as a third dimension in defining the ordering, but it is not obvious that this would be the most promising approach.

1.2.5 Spatial kernel stick-breaking prior

An extension of the stick-breaking prior of [18] to the multivariate spatial setting is proposed in [28]. The stick-breaking prior can be extended to the univariate spatial setting by incorporating spatial information into either the model for the locations $\boldsymbol{\theta}_i$ or the model for the weights p_i. As explained in Subsection 1.2.2, [23] models the locations as vectors drawn from a spatial distribution. This approach is generalized by [24] to allow both the weights and locations to vary spatially. However, we have seen that these models require replication of the spatial process. As discussed in the previous subsection, [27] proposes a spatial Dirichlet model that does not require replication. The latter model permutes the V_i

based on spatial location, allowing the occurrence of $\boldsymbol{\theta}_i$ to be more or less likely in different regions of the spatial domain. The nonparametric multivariate spatial model introduced by [28] has multivariate normal priors for the locations $\boldsymbol{\theta}_i$. We call this prior process a spatial kernel stick-breaking (SSB) prior. Similar to [27], the weights p_i vary spatially. However, rather than random permutation of V_i, [28] introduces a series of kernel functions to allow the masses to change with space. This results in a flexible spatial model, as different kernel functions lead to different relationships between the distributions at nearby locations. This model is similar to that of [29], who use kernels to smooth the weights in the non-spatial setting. This model is also computationally convenient because it avoids reversible jump MCMC steps and the inversion of large matrices.

In this section, first, we introduce the SSB prior in the univariate setting, and then we extend it to the multivariate case. Let $Y(\mathbf{s})$, the observable value at site $\mathbf{s} = (s_1, s_2)$, be modelled as

$$Y(\mathbf{s}) = \eta(\mathbf{s}) + \mathrm{x}(\mathbf{s})^T\boldsymbol{\beta} + \epsilon(\mathbf{s}), \tag{1.14}$$

where $\eta(\mathbf{s})$ is a spatial random effect, $\mathrm{x}(\mathbf{s})$ is a vector of covariates for site $\mathbf{s}$, $\boldsymbol{\beta}$ are the regression parameters, and $\epsilon(\mathbf{s}) \overset{iid}{\sim} N(0, \tau^2)$.

The spatial effects are assigned a random prior distribution, i.e., $\eta(\mathbf{s}) \sim F_{\mathbf{s}}(\eta)$. This SSB modelling framework introduces models marginally, i.e., $F_{\mathbf{s}}(\eta)$ and $F_{\mathbf{s}'}(\eta)$, rather than jointly, i.e. $F_{\mathbf{s},\mathbf{s}'}(\eta)$ as in the referenced work of Gelfand and colleagues. The distributions $F_{\mathbf{s}}(\eta)$ are smoothed spatially. Extending (1.10) to depend on $\mathbf{s}$, the prior for $F_{\mathbf{s}}(\eta)$ is the potentially infinite mixture

$$F_{\mathbf{s}}(\eta) \overset{d}{=} \sum_{i=1}^{N} p_i(\mathbf{s})\delta_{\boldsymbol{\theta}_i}, \tag{1.15}$$

where $p_i(\mathbf{s}) = V_i(\mathbf{s})\prod_{j=1}^{i-1}(1 - V_j(\mathbf{s}))$, and $V_i(\mathbf{s}) = w_i(\mathbf{s})V_i$. The distributions $F_{\mathbf{s}}(\eta)$ are related through their dependence on the V_i and $\boldsymbol{\theta}_i$, which are given the priors $V_i \sim \mathrm{Beta}(a, b)$ and $\boldsymbol{\theta}_i \sim H$, each independent across i. However, the distributions vary spatially according to the functions $w_i(\mathbf{s})$, which are restricted to the interval $[0, 1]$. $w_i(\mathbf{s})$ is modelled using a kernel function, but alternatively $\log(w_i(\mathbf{s})/(1 - w_i(\mathbf{s})))$ could be modelled as a spatial Gaussian process. Other transformations could be considered, but we use kernels for simplicity. There are many possible kernel functions and Table 1.4 gives three examples. In each case, the func-

Table 1.4: Examples of kernel functions and the induced functions $\gamma(\mathbf{s}, \mathbf{s}')$, where $\mathbf{s} = (s_1, s_2)$, $h_1 = |s_1 - s_1'| + |s_2 - s_2'|$, $h_2 = \sqrt{(s_1 - s_1')^2 + (s_2 - s_2')^2}$, $I(\cdot)$ is the indicator function, and $x^+ = \max(x, 0)$.

Name	$w_i(\mathbf{s})$	Model for κ_{1i} and κ_{2i}	$\gamma(\mathbf{s}, \mathbf{s}')$
Uniform	$\prod_{j=1}^2 I(\lvert s_j - \psi_{ji}\rvert < \frac{\kappa_{ji}}{2})$	$\kappa_{1i}, \kappa_{2i} \equiv \lambda$	$\prod_{j=1}^2 \left(1 - \frac{\lvert s_j - s_j'\rvert}{\lambda}\right)^+$
Uniform	$\prod_{j=1}^2 I(\lvert s_j - \psi_{ji}\rvert \frac{<}{2} \kappa_{ji})$	$\kappa_{1i}, \kappa_{2i} \sim$Exp($\lambda$)	$\exp(-h_1/\lambda)$
Exponential	$\prod_{j=1}^2 \exp(-\frac{\lvert s_j - \psi_{ji}\rvert}{\kappa_{ji}})$	$\kappa_{1i}, \kappa_{2i} \equiv \lambda$	$0.25\left[\prod_{j=1}^2 \left(1 + \frac{\lvert s_j - s_j'\rvert}{\lambda}\right)\right] \exp\left(-\frac{h_1}{\lambda}\right)$
Squared exp.	$\prod_{j=1}^2 \exp(-\frac{(s_j - \psi_{ji})^2}{\kappa_{ji}^2})$	$\kappa_{1i}, \kappa_{2i} \equiv \lambda^2/2$	$0.5 \exp\left(-\frac{h_2^2}{\lambda^2}\right)$
Squared exp.	$\prod_{j=1}^2 \exp(-\frac{(s_j - \psi_{ji})^2}{\kappa_{ji}^2})$	$\kappa_{1i}, \kappa_{2i} \sim$InvGa($\frac{3}{2}, \frac{\lambda^2}{2}$)	$0.5/\left(1 + (\frac{h_2}{\lambda})^2\right)$

tion $w_i(\mathbf{s})$ is centered at knot $\boldsymbol{\psi}_i = (\psi_{1i}, \psi_{2i})$ and the spread is controlled by the bandwidth parameter $\boldsymbol{\kappa}_i = (\kappa_{1i}, \kappa_{2i})$. Both the knots and the bandwidths are modelled as unknown parameters. The knots $\boldsymbol{\psi}_i$ are given independent uniform priors over the spatial domain. The bandwidths can be modelled as equal for each kernel function or varying independently following distributions given in the third column of Table 1.4.

To ensure that the stick-breaking prior is proper, we must choose priors for κ_i and V_i so that $\sum_{i=i}^N p_i(\mathbf{s}) = 1$ almost surely for all $\mathbf{s}$. [28] show that the SSB prior with infinite N is proper if $E(V_i) = a/(a+b)$ and $E\left[w_i(\mathbf{s})\right]$ (where the expectation is taken over $(\boldsymbol{\psi}_i, \boldsymbol{\kappa}_i)$) are both positive. For finite N, we can ensure that $\sum_{i=i}^N p_i(\mathbf{s}) = 1$ for all $\mathbf{s}$ by setting $V_N(\mathbf{s}) \equiv 1$ for all $\mathbf{s}$. This is equivalent to truncating the infinite mixture by attributing all of the mass from the terms with $i \geq N$ to $p_N(\mathbf{s})$.

In practice, allowing N to be infinite is often unnecessary and computationally infeasible. Choosing the number of components in a mixture model is notoriously problematic. Fortunately, in this setting the truncation error can easily be assessed by inspecting the distribution of $p_N(\mathbf{s})$, the mass of the final component of the mixture. The number of components N can be chosen by generating samples from the prior distribution of $p_N(\mathbf{s})$. We increase N until $p_N(\mathbf{s})$ is satisfactorily small for each site $\mathbf{s}$. Also, the truncation error is monitored by inspecting the posterior distribution of $p_N(\mathbf{s})$, which is readily available from

the MCMC samples.

Assuming a finite mixture, the spatial stick-breaking model can be written as a finite mixture model where $g(\mathbf{s}) \in \{1, ..., N\}$ indicates the particular location allocated to site $\mathbf{s}$, i.e,

$$\begin{aligned} Y(\mathbf{s}) &= \theta_{g(\mathbf{s})} + \mathrm{x}(\mathbf{s})^T\boldsymbol{\beta} + \epsilon(\mathbf{s}), \text{ where } \epsilon(\mathbf{s}) \overset{iid}{\sim} N(0, \tau^2) \\ \theta_j &\overset{iid}{\sim} N(0, \sigma^2), j = 1, ..., N \\ g(\mathbf{s}) &\sim \text{Categorical}(p_1(\mathbf{s}), ..., p_N(\mathbf{s})) \\ p_j(\mathbf{s}) &= w_j(\mathbf{s})V_j \prod_{k<j} \left[1 - w_k(\mathbf{s})V_k\right], \text{ where } V_j \overset{iid}{\sim} \text{Beta}(a, b) \end{aligned} \tag{1.16}$$

where $\eta(\mathbf{s}) = \theta_{g(\mathbf{s})}$. The regression parameters $\boldsymbol{\beta}$ are given vague normal priors. This model can also easily be fitted if we have an infinite mixture model with $N = \infty$, using retrospective sampling ideas from [21].

Understanding the spatial correlation function is crucial for analyzing spatial data. Although the SSB prior foregoes the Gaussian assumption for the spatial random effects, we can still compute and investigate the covariance function. Conditional on the probabilities $p_j(s)$ (but not the locations $\boldsymbol{\theta}_j$), the covariance between two observations is

$$\text{cov}(Y(\mathbf{s}), Y(\mathbf{s}')) = \sigma^2 P\left(\eta(\mathbf{s}) = \eta(\mathbf{s}')\right) = \sigma^2 \sum_{j=1}^{N} p_j(\mathbf{s})p_j(\mathbf{s}'). \tag{1.17}$$

For a one-dimensional spatial domain, integrating over (V_i, ψ_i, κ_i) and letting $N \to \infty$ gives

$$\text{Var}\left(Y(s)\right) = \sigma^2 + \tau^2 \tag{1.18}$$

$$\text{Cov}(Y(s), Y(s')) = \sigma^2\gamma(s, s')\left[2\frac{a+b+1}{a+1} - \gamma(s, s')\right]^{-1}, \tag{1.19}$$

where

$$\gamma(s, s') = \frac{\int\int w_i(s)w_i(s')p(\psi_i, \kappa_i)d\psi_i d\kappa_i}{\int\int w_i(s)p(\psi_i, \kappa_i)d\psi_i d\kappa_i} \in [0, 1]. \tag{1.20}$$

Since (V_i, ψ_i, κ_i) have independent priors that are uniform over the spatial domain, integrating over these parameters gives a stationary prior covariance. However, the conditional

covariance can be non-stationary. More importantly, the posterior predictive distribution can accommodate non-stationarity. Therefore, we conjecture the SSB model is more robust to non-stationarity than traditional stationary Kriging methods.

If $b/(a+1)$ is large, i.e., the V_i are generally small and there are many terms in the mixture with significant mass, the correlation between $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ is approximately proportional to $\gamma(\mathbf{s}, \mathbf{s}')$. Table 1.4 from [28] gives the function $\gamma(\mathbf{s}, \mathbf{s}')$ for several examples of kernel functions and shows that different kernels can produce very different correlation functions.

To introduce a multivariate extension of the SBB prior, let $\boldsymbol{\eta}(\mathbf{s}) = (\eta_1(\mathbf{s}), \ldots, \eta_p(\mathbf{s}))^T$ be a $p-$dimensional spatial process. The prior for $\boldsymbol{\eta}(\mathbf{s})$ is

$$\boldsymbol{\eta}(\mathbf{s}) \sim F_{\mathbf{s}}(\boldsymbol{\eta}), \text{ where } F_{\mathbf{s}}(\boldsymbol{\eta}) \stackrel{d}{=} \sum_{i=1}^{N} p_i(\mathbf{s}) \delta_{\boldsymbol{\theta}_i}. \tag{1.21}$$

The weights $p_i(\mathbf{s})$ are shared across components, the p-dimensional locations $\boldsymbol{\theta}_i \stackrel{iid}{\sim} N(0, \Sigma)$, and Σ is a $p \times p$ covariance matrix that controls the association between the p spatial processes. The covariance Σ could have an Inverse-Wishart prior.

1.2.6 A case study: Hurricane Ivan

In [28] the multivariate SSB prior is used to model the complex spatial patterns of hurricane wind vectors, with data from Hurricane Ivan as it passes through the Gulf of Mexico at 12 pm on September 15, 2004. Three sources of information are used in the analysis and are plotted in Figure 1.4. The first source is gridded satellite data (Figure 1.4a) available from NASA's SeaWinds database (`http://podaac.jpl.nasa.gov/products/product109.html`). These data are available twice daily on a 0.25 x 0.25 degree, global grid. Due to the satellite data's potential bias, measurement error, and coarse temporal resolution, the wind fields analysis is supplemented with data from the National Data Buoy Center of the National Oceanic and Atmospheric Administration (NOAA). Buoy data are collected every ten minutes at a relatively small number of marine locations (Figure 1.4b). These measurements are adjusted to a common height of 10 meters above sea level using the algorithm of [34].

In addition to satellite and buoy data, the deterministic Holland model [30] is incorporated in the analysis. The NOAA currently uses this model alone to produce wind fields for their numerical ocean models. The Holland model predicts that the wind velocity at location $\mathbf{s}$ is

$$H(\mathbf{s}) = \left(\frac{B}{\rho} \left(\frac{Rmax}{r} \right)^B (P_n - P_c) \exp \left[- \left(\frac{Rmax}{r} \right)^B \right] \right)^{1/2}, \tag{1.22}$$

where r is the radius from the storm center to site $\mathbf{s}$, P_n is the ambient pressure, P_c is the hurricane central pressure, ρ is the air density, $Rmax$ is the radius of maximum sustained winds and B controls the shape of the pressure profile.

We decompose the wind vectors into their orthogonal west/east (u) and north/south (v) vectors. The Holland model for the u and v components is

$$H_u(\mathbf{s}) = H(\mathbf{s}) sin(\phi) \quad \text{and} \quad H_v(\mathbf{s}) = H(\mathbf{s}) cos(\phi), \tag{1.23}$$

where ϕ is the inflow angle at site $\mathbf{s}$, across circular isobars toward the storm center, rotated to adjust for the storm's direction. We fix the parameters $P_n = 1010$mb, $P_c = 939$mb, $\rho = 1.2$ kg m^{-3}, and $Rmax = 49$, and $B = 1.9$ using the meteorological data from the national hurricane center (`http://www.nhc.noaa.gov`) and recommendations of [31]. The output from this model for Hurricane Ivan is plotted in Figure 1.4c. By construction, Holland model output is symmetric with respect to the storm's center, which does not agree with the satellite observations in Figure 1.4a.

Let $u(\mathbf{s})$ and $v(\mathbf{s})$ be the underlying wind speed in the west/east and north/south directions, respectively, for spatial location $\mathbf{s}$. We distinguish the different sources of wind data: $u_T(\mathbf{s})$ and $v_T(\mathbf{s})$ are satellite measurements and $u_B(\mathbf{s})$ and $v_B(\mathbf{s})$ are buoy measurements. The model used by [28] for these data is

$$\begin{aligned} u_T(\mathbf{s}) &= a_u + u(\mathbf{s}) + e_{uT}(\mathbf{s}) \qquad & v_T(\mathbf{s}) &= a_v + v(\mathbf{s}) + e_{vT}(\mathbf{s}) \\ u_B(\mathbf{s}) &= u(\mathbf{s}) + e_{uB}(\mathbf{s}) \qquad & v_B(\mathbf{s}) &= v(\mathbf{s}) + e_{vB}(\mathbf{s}), \end{aligned} \tag{1.24}$$

where $\{e_{uT}, e_{vT}, e_{uB}, e_{vB}\}$ are independent (from each other and from the underlying winds), zero mean, Gaussian errors, each with its own variance, and $\{a_u, a_v\}$ account for additive

bias in the satellite and aircraft data. Of course, the buoy data may also have bias, but it is impossible to identify bias from both sources, so all the bias is attributed to the satellite measurements.

The underlying orthogonal wind components $u(\mathbf{s})$ and $v(\mathbf{s})$ are modelled as a mixture of a deterministic wind model and a semiparametric multivariate spatial process

$$u(\mathbf{s}) = H_u(\mathbf{s}) + R_u(\mathbf{s}) \tag{1.25}$$

$$v(\mathbf{s}) = H_v(\mathbf{s}) + R_v(\mathbf{s}), \tag{1.26}$$

where $H_u(\mathbf{s})$ and $H_v(\mathbf{s})$ are the orthogonal components of the deterministic Holland model in (1.23) and $\mathbf{R}(\mathbf{s}) = (R_u(\mathbf{s}), R_v(\mathbf{s}))'$ follows a multivariate extension of the SSB prior as in (1.21). The prior for $\mathbf{R}(\mathbf{s})$ is

$$\mathbf{R}(\mathbf{s}) \sim F_{\mathbf{s}}(\eta), \text{ where } F_{\mathbf{s}}(\eta) \stackrel{d}{=} \sum_{i=1}^{N} p_i(\mathbf{s})\delta_{\boldsymbol{\theta}_i}. \tag{1.27}$$

The masses $p_i(\mathbf{s})$ are shared across components, the two-dimensional locations $\boldsymbol{\theta}_i \stackrel{iid}{\sim} N(0, \Sigma)$, and Σ is a 2×2 covariance matrix that controls the association between the two wind components. The inverse covariance Σ^{-1} has a Wishart prior with 3.1 degrees of freedom and inverse scale matrix $0.1\,\mathbf{I}_2$. After transforming the spatial grid to be contained in the unit square, the spatial knots ψ_{1i} and ψ_{2i} have independent Beta(1.5,1.5) priors to encourage knots to lie near the center of the hurricane where the wind is most volatile.

In [28], this multivariate SSB model is fitted to 182 satellite observations and 7 buoy observations for Hurricane Ivan. To illustrate the effect of relaxing the normality assumption, [28] also fits a fully-Gaussian model that replaces the stick-breaking prior for $\mathbf{R}(\mathbf{s})$ in (1.27) with a zero-mean Gaussian prior

$$\text{Var}(\mathbf{R}(\mathbf{s})) = \Sigma \quad \text{and} \quad \text{Cov}(\mathbf{R}(\mathbf{s}), \mathbf{R}(\mathbf{s}')) = \Sigma \times \exp\left(-||\mathbf{s} - \mathbf{s}'||/\lambda\right), \tag{1.28}$$

where Σ controls the dependency between the wind components at a given location and λ is a spatial range parameter. The covariance parameters Σ and λ have the same priors as the

covariance parameters in the SSB model.

Since the primary objective is to predict wind vectors at unmeasured locations to use as inputs for numerical ocean models, statistical models are compared in terms of expected mean squared prediction error ([32]; [33]). The EMSPE is smaller for the semiparametric model with uniform kernels (EMSPE=3.46) than for the semiparametric model using squared exponential kernels (EMSPE=4.19) and the fully Gaussian model (EMSPE=5.17).

Figure 1.5 summarizes the posterior from the SSB prior with uniform kernel functions. The fitted values in Figures 1.5a and 1.5b vary rapidly near the center of the storm and are fairly smooth in the periphery. After accounting for the Holland model, the correlation between the residual u and v components $R_u(\mathbf{s})$ and $R_v(\mathbf{s})$ ($\Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$, where Σ_{kl} is the (k,l) element of Σ) is generally negative, confirming the need for a multivariate analysis.

To show that the multivariate SSB model with uniform kernel functions fits the data well, 10% of the observations are randomly set aside (across u and v components and buoy and satellite data) throughout the model fitting and the 95% predictive intervals for the missing observations are obtained. The prediction intervals contain 94.7% (18/19) of the deleted u components and 95.2% (20/21) of the deleted v components. These statistics suggest that the model is well-calibrated.

References

[1] De Oliveira, V., Kedem, B. and Short, D.A. (1997). "Bayesian prediction of transformed Gaussian random fields," *Journal of the American Statistical Association*, **92**, 1422–1433.

[2] Diggle, P.J. Tawn, J.A. and Moyeed, R.A. (1998). "Model-based geostatistics (with discussion)," *Applied Statistics*, **47**, 299–326.

[3] Banerjee, S., Wall, M. and Carlin, B.P. (2003). "Frailty modeling for spatially correlated survival data with application to infant mortality in Minnesota," *Biostatistics*, **4**, 123–142.

[4] Li, Y. and Ryan, L. (2002). "Modeling Spatial Survival Data Using Semiparametric

Frailty Models," *Biometrics*, **58**, 287–297.

[5] Brown, P.E., Diggle, P.J. and Henderson, R. (2003). "A non-Gaussian spatial process model for opacity of flocculated paper," *Scandinavian Journal of Statistics*, **30**, 355–368.

[6] Palacios, M.B. and Steel, M.F.J. (2006). "Non-Gaussian Bayesian geostatistical modelling," *Journal of the American Statistical Association*, **101**, 604–618.

[7] Damian, D., Sampson, P.D. and Guttorp, P. (2001). "Bayesian Estimation of Semi-parametric Non-stationary Spatial Covariance Structures," *Environmetrics*, **12**, 161–178.

[8] Damian, D., Sampson P.D. and Guttorp, P. (2003). "Variance modeling for nonstationary processes with temporal replications," *Journal of Geophysical Research (Atmosphere)*, **108**, 8778.

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

[10] Verdinelli, I. and Wasserman, L. (1995). "Computing Bayes factors by using a generalization of the Savage-Dickey density ratio," *Journal of the American Statistical Association*, **90**, 614–618.

[11] Berger, J.O., De Oliveira, V. and Sansó, B. (2001). "Objective Bayesian analysis of spatially correlated data," *Journal of the American Statistical Association*, **96**, 1361–1374.

[12] Stein, M.L. (1999). *Interpolation of Spatial Data. Some Theory of Kriging.* Springer-Verlag, New York.

[13] Müller, P. and Quintana, F.A. (2004). "Nonparametric Bayesian data analysis," *Statistical Science*, 95–110.

[14] Dunson, D.B. (2008). "Nonparametric Bayes applications to biostatistics," in *Nonparametric Bayesian Statistics*, Cambridge: Cambridge Univ. Press, to appear.

[15] Schmidt, A.M. and O'Hagan, A. (2003). "Bayesian inference for nonstationary spatial covariance structure via spatial deformations," *Journal of the Royal Statistical Society*, B, **65**, 745–758.

[16] Ishwaran, H. and James, L. (2001). "Gibbs-sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, **96**, 161–173.

[17] Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, **1**, 209–30.

[18] Sethuraman, J. (1994). "A constructive definition of Dirichlet priors," *Statistica Sinica*, **4**, 639–50.

[19] Antoniak, C.E. (1974). "Mixtures of Dirichlet processes with applications to nonparametric problems," *Annals of Statistics*, **2**, 1152–1174.

[20] MacEachern, S.N. (1998). "Computational methods for mixture of Dirichlet process models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, D. Dey, P. Müller and D. Sinha, eds. New York: Springer-Verlag, pp. 23–44.

[21] Papaspiliopoulos O, Roberts G (2008). "Retrospective MCMC for Dirichlet process hierarchical models," *Biometrika*, **95**, 169-186.

[22] MacEachern, S.N. (1999). "Dependent nonparametric processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association.

[23] Gelfand, A.E., Kottas, A. and MacEachern, S.N. (2005). "Bayesian nonparametric spatial modelling with Dirichlet processes mixing", *Journal of the American Statistical Association*, **100**, 1021–1035.

[24] Duan, J.A., Guindani, M. and Gelfand, A.E. (2007). "Generalized spatial Dirichlet process models", *Biometrika*, **94**, 809–825.

[25] Gelfand, A.E., Guindani, M. and Petrone, S. (2007). "Bayesian nonparametric modeling for spatial data analysis using Dirichlet processes," in *Bayesian Statistics,* **8** J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, eds., Oxford: Oxford Univ. Press.

[26] Petrone, S., Guindani, M. and Gelfand, A.E. (2008). "Hybrid Dirichlet mixture models for functional data", mimeo. **Status?**

[27] Griffin, J.E. and Steel, M.F.J. (2006). "Order-based dependent Dirichlet processes," *Journal of the American Statistical Association*, **101**, 179–194.

[28] Reich, B., and Fuentes, M. (2007). "A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields," *Annals of Applied Statistics*, **1**, 249-264.

[29] Dunson, D.B. and Park, J. H. (2008). "Kernel stick-breaking processes," *Biometrika*,

95, 307-323.

[30] Holland, G.J. (1980). "An analytic model of the wind and pressure profiles in hurricanes." *Monthly Weather Review*, **108**, 1212-1218.

[31] Hsu, S.A. and Yan, Z. (1998). "A note on the radius of maximum wind for hurricanes." *Journal of Coastal Research*, **14**, 667–668.

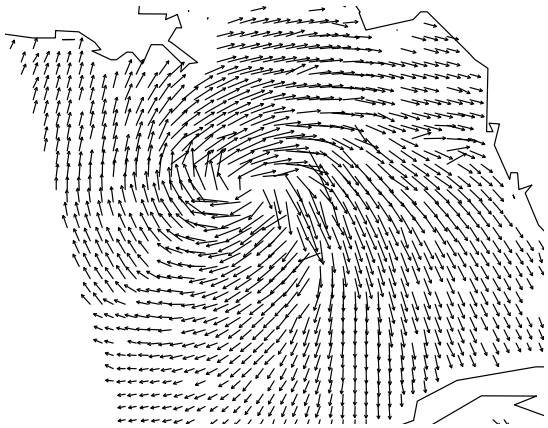
[32] Laud, P. and Ibrahim, J. (1995). "Predictive model selection." *Journal of the Royal Statistical Society, Series B*, **57**, 247–262.

[33] Gelfand, A.E. and Ghosh, S.K. (1998). "Model choice: a minimum posterior predictive loss approach." *Biometrika*, **77**, 1–11.

[34] Large, W.G. and Pond, S. (1981). "Open ocean momentum flux measurements in moderate to strong winds." *Journal of Physical Oceanography*, **11**, 324–336.

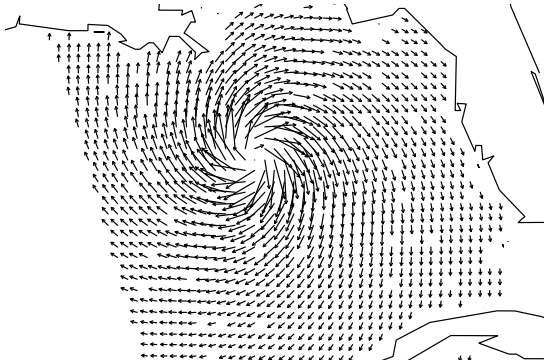
Figure 1.4: Plot of various types of wind field data/output for Hurricane Ivan on September 15, 2004.

(a) Satellite data



(b) Buoy data

(c) Holland model output



(d) Satellite - Holland model output

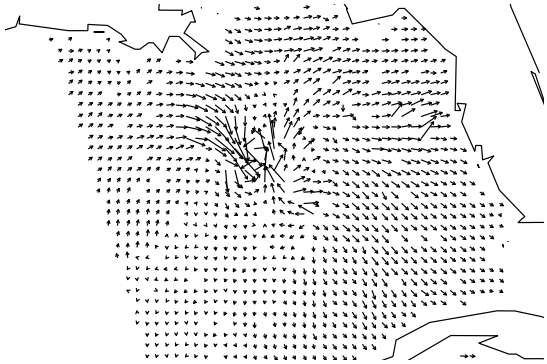


Figure 1.5: Summary of the posterior of the spatial stick-breaking model with uniform kernels. Panels (a) and (b) give the posterior mean surface for the u and v components.