BOHS
The Chartered Society for
Worker Health Protection

OXFORD

Original Article

Coastline Kriging: A Bayesian Approach

Nada Abdalla[1], Sudipto Banerjee[1,*], Gurumurthy Ramachandran[2], Mark Stenzel[3] and Patricia A. Stewart[3]

[1]Department of Biostatistics, University of California-Los Angeles, 650 Charles E. Young Drive South, Los Angeles, CA 91302, USA; [2]Department of Environmental Health and Engineering, Bloomberg School of Public Health, Johns Hopkins University, 615 North Wolfe St, Baltimore, MD 21205, USA; [3]Stewart Exposure Assessments, LLC, 6045 North 27th St, Arlington, VA 22207, USA

*Author to whom correspondence should be addressed. Tel: +310-825-5916; E-mail: sudipto@ucla.edu

Abstract

Statistical interpolation of chemical concentrations at new locations is an important step in assessing a worker's exposure level. When measurements are available from coastlines, as is the case in coastal clean-up operations in oil spills, one may need a mechanism to carry out spatial interpolation at new locations along the coast. In this article, we present a simple model for analyzing spatial data that is observed over a coastline. We demonstrate four different models using two different representations of the coast using curves. The four models were demonstrated on simulated data and one of them was also demonstrated on a dataset from the GuLF STUDY (Gulf Long-term Follow-up Study). Our contribution here is to offer practicing hygienists and exposure assessors with a simple and easy method to implement Bayesian hierarchical models for analyzing and interpolating coastal chemical concentrations.

Keywords: coastal kriging; Gaussian process; geostatistics; hierarchical modeling; kriging; Markov chain Monte Carlo

Introduction

Data observed over locations with known geographic coordinates are often referred to as point-referenced data and are commonly seen in environmental health. Recent applications consider such data measured along coastlines or shores. For example, assessing exposures of workers to chemicals along the coastline may require statistical interpolation of the chemical concentration at unmonitored locations along the coast. Statistical interpolation at new locations based upon a set of observed measurements at known locations is often referred to as 'kriging' in the geostatistical literature (Cressie, 1993). Kriging customarily uses spatial analytic tools such as variograms or covariance functions to construct best linear unbiased predictors. When chemicals are sampled mostly along a coastline, interpolation is sought at new locations along the coast. Thus, all measurements are collected along a curve (approximating the coastline) and prediction is sought at new points on this curve. We call this 'coastal kriging'.

Models for waterway stream networks using moving averages have been developed (Hoef and Peterson, 2010). They use stream distance rather than Euclidean distance. These models account for the volume and direction of flowing water in stream networks. They offer richness and flexibility but are complicated and can be difficult to compute. Unlike networks, where we have complex structure of line segments and joints, in simple coastal kriging, we approximate the coastline with a single curve or a sequence of line segments. A simple parametrization of the coast will suffice and lead to easily implementable statistical models.

We will pursue Bayesian coastal kriging. Bayesian models offer easier interpretability for parameter estimates, provide exact estimates of uncertainty without requiring assumptions of large sample sizes and independence of observations, and can incorporate prior information when available. Incorporating prior information is not uncommon in exposure assessment and can improve decision making (e.g. Ramachandran *et al.*, 2003; Hewett *et al.*, 2006). Bayesian models can be easily executed using several software packages within the R statistical computing environment (please refer to the coding material as seen in the Supplementary Material, available at *Annals of Occupational Hygiene* online).

We will illustrate our models using a specific dataset extracted from the GuLF STUDY (Gulf Long-term Follow-up Study) database. In April 2010, an explosion of the Deepwater Horizon oil rig resulted in an oil spill in the Gulf of Mexico. It was the largest oil spill in the US history. Tens of thousands of workers were involved in stopping and in cleaning up the oil release. The GuLF STUDY is conducted by the National Institute of Environmental Health Sciences (NIEHS) and sponsored by the National Institute of Health (NIH) (Kwok *et al.*, 2018). It is collecting information to study potential adverse effects on the health of those workers. Among other activities, the workers capped the well, applied dispersants to break up the oil, skimmed or burned the oil on the Gulf waters, cleaned beaches, marshes and structures, decontaminated equipment, and provided support for these activities. Personal air measurements are available on many of these tasks. The highest portion of the STUDY participants were involved in cleaning the beaches, marshes, and structures. One specific task in assessing exposures of workers cleaning the coastline is to statistically interpolate the chemical concentration at new locations along the coast.

Our contribution in this article expands upon existing geostatistical models to allow for better prediction of quantities of interest at new locations over coastlines. The article is organized as follows.

The next section ("Model-based kriging") provides a brief review of Bayesian methods for kriging. Next we discuss "Spatial processes for coastline measurements". The section titled "Coastal Kriging" discusses our geostatistical models for interpolating point-referenced coastline data and simple algorithms for implementing Bayesian kriging. This is followed by a section on simulation experiments that help validate our method. "Data Analysis" illustrates our model through applying it to the GuLF STUDY data. Finally, we conclude the article suggesting some future work.

Model-based Kriging

Spatial process models with Euclidean coordinates

Point-referenced spatial modeling seeks to capture associations between observations geographically closer to each other and to predict the value of the response or outcome variable at arbitrary locations. This is achieved using a spatial regression model:

$$Z(s) = x(s)^{\mathrm{T}} \beta + \omega(s) + \varepsilon(s), \, \varepsilon(s) \overset{iid}{\sim} N\left(0, \tau^2\right), \quad (1)$$

where $x(s)^{\mathrm{T}}$ is a $1 \times p$ vector of covariates (predictors) observed at location s, $\omega(s)$ is a latent (unobserved) spatial random effect at location s, and $\varepsilon(s)$ accounts for measurement error. For any collections of locations, the measurement errors in equation (1) are normally distributed independently and identically, each with a zero mean and variance τ^2.

If $\omega(s) = 0$ for all locations, then equation (1) reduces to an ordinary linear model with independent outcomes. If the outcomes are spatially correlated, then $\omega(s)$ introduces dependence. There are several different mechanisms for specifying $\omega(s)$ (Cressie, 1993; Banerjee *et al.*, 2014), but we choose a fairly straightforward and interpretable model here. We assume that each $\omega(s)$ has mean 0 and the dependence at two points s and s' is modeled as follows:

$$\mathrm{Cov}\{\omega(s), \omega(s')\} = K_\theta\left(s, s'\right) = \sigma^2 \exp\left(-\phi \| s - s' \|\right), \quad (2)$$

where $\| s - s' \|$ is the distance between two locations s and s', σ^2 captures the variation attributed to spatial effects (referred to as partial sill) and ϕ controls the rate at which the spatial correlation drops to zero. The process parameters $\theta = \{\sigma^2, \phi\}$ together with the distance between any two points completely specify the spatial covariance function $K_\theta(s, s')$ The *spatial range* is defined as the distance beyond which the spatial correlation becomes negligible. For the exponential covariance

function in equation (2), the spatial range is given by approximately $3/\phi$ which is the distance where the correlation drops <0.05.

We incorporate the covariance function (equation 2) into a probability model. Let $S = \{s_1, s_2, ..., s_n\}$ be the set of spatial locations. The $n \times 1$ vector ω, whose ith entry is $\omega(s_i)$, follows a multivariate normal distribution $N(0, K_\theta)$, where K_θ is the $n \times n$ spatial covariance matrix with (i, j)th entry $K_\theta(s_i, s_j)$ in equation (2). The measurement errors are independent across locations, hence $\varepsilon(s_i) \overset{iid}{\sim} N(0, \tau^2)$. This implies that the data vector Z, whose ith element is $Z(s_i)$, is multivariate normal with mean vector $X\beta$, where $x(s_i)^{\mathrm{T}}$ are the rows of X, and variance–covariance matrix $K_\theta + \tau^2 I_n$, where I_n denotes the $n \times n$ identity matrix.

Spatial regression models, such as equation (1), are fitted by estimating geostatistical parameters σ^2, ϕ, and τ^2 in addition to the regression coefficients β. We use equation (1) to predict the outcome at a new location after accounting for the uncertainty in parameter estimates. When all points lie on a region represented as a 2-D plane, the distance between s and s' in equation (2) is given by the standard Euclidean distance formula. Here, the correlation drops at the same rate for every direction, so the spatial range is a function of distance only. Also, the covariance function in equation (2) ensures that K_θ is always positive definite (e.g. Banerjee *et al.*, 2014).

In our current context, the points lie along a curve representing the coastline. There are two issues. First, the Euclidean distance is inappropriate for modeling spatial covariances because the effective spatial range will be the distance *along the coast* at which the correlation becomes negligible. Second, covariance functions that ensure positive definiteness in Euclidean coordinates need not be valid for other domains (Banerjee, 2005). This means that we will need to construct valid covariance functions along the coastline. Subsequently, we describe a simple approach to construct models such as equation (1) using valid covariance functions for points along curves.

Spatial processes for coastline measurements

We now extend the model discussed in the previous section to the case where the data are observed over a coastline. Since all observations lie along the coastline, we will model spatial dependence along the coastline. The spatial range and variability will need to be interpreted in terms of distance along the coastline. Prediction is also sought at arbitrary points along the coast. We assume that any point s on the coast is given by $\gamma(t) = (\gamma_1(t), \gamma_2(t))$ for some $t \in T \subset \Re^1$, where $\gamma_1(t) = f(t)$ and $\gamma_2(t) = g(t)$ are parametric equations for the coordinates. Therefore, each value of t determines a coordinate on a

plane and traces out a curve $\gamma(t)$ as t varies over a range T. The coastline is now given by the set of all points on it: $\gamma(T) = \{\gamma(t) : t \in T \subset \Re^1\}$. For example, a simple curve could be approximated by line segments. For each line segment, $\gamma(t)$ is a straight line, $\gamma_i(t) = \{s_i + tu | t \in [0, \infty]\}$, originating at s_i and parallel to the direction vector u.

Here, $s_i = \begin{bmatrix} \gamma_{1i} \\ \gamma_{2i} \end{bmatrix}$, $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and, hence,

$$\gamma_i(t) = \left\{ \begin{bmatrix} \gamma_{1i} + tu_1 \\ \gamma_{2i} + tu_2 \end{bmatrix} \middle| t \in [0, \infty] \right\}. \tag{3}$$

A customary choice for the parameter t is the arc-length. As another example, consider a circular coast with radius r. The curve is defined as

$$\gamma(t) = \{\gamma_1(t) = r\cos t, \gamma_2(t) = r\sin t | t \in [0, \pi/2]\}. \tag{4}$$

The point $\gamma(t) = (r \cos t, r \sin t)$ moves in a fixed orientation (e.g. clockwise) as t increases. If t is the length of an arc of the circle and λ is the angle in radians which the arc subtends at the center of the circle, then $t = r\lambda$.

A spatial regression model such as equation (1) can be defined over a coast by representing each point on the coast by $\gamma(t)$. Thus, we write $Y(t) = Z(\gamma(t))$ for every $t \in T$. Therefore,

$$Y(t) = x^{\mathrm{T}}(t)\beta + \omega(t) + \varepsilon(t), \tag{5}$$

where $x(t)$ is the vector of covariates observed at the point $\gamma(t)$, $\omega(t)$ is now defined over T with covariance function

$$\mathrm{Cov}(\omega(t), \omega(t')) = K_\theta(t, t') = \sigma^2 \exp(-\phi | t - t' |), \tag{6}$$

where $| t - t' |$ is the absolute difference between t and t', and $\varepsilon(t) \overset{iid}{\sim} N(0, \tau^2)$.

The choice of t depends on the parametric equation used to approximate the coast. If the coast can be well represented in closed form using a parametric equation, then t as the arc-length is often a reasonable and convenient choice (e.g. Stewart, 2012). The intuition stems from describing a curve by starting at a point on the curve and moving along the path traced by that curve. The point we arrive at after moving a distance of t units on the curve is $\gamma(t)$. More generally, an arbitrary coastline can be well approximated using a series of small line segments. Each segment is then defined according to equation (3). For example, in our subsequent simulation experiments, we present linear approximations for an elliptical coastline. In our real example, we use a series of small linear segments to model the coast along Waveland Beach in Mississippi.

Coastal kriging

Exposure assessors may be interested in predicting the concentration of a toxicant at any arbitrary location on the coast. Let $Y(t_0)$ be the toxicant concentration measurement at the point $\gamma(t_0)$ on the coast. The posterior probability distribution of $Y(t_0)$, which is also referred to as the posterior predictive distribution, is computed in two steps. First, the unknown parameters in equation (5) are estimated by using Bayes' theorem to compute their posterior distributions. Thus, if $p(\theta, \beta, \tau^2)$ represents the prior distribution for unknown parameters and $p(y|\theta, \beta, \tau^2)$ represents the likelihood, then the posterior distribution is given by

$$p(\theta,\beta,\tau^2|y) = \frac{p(\theta,\beta,\tau^2) \times p(y|\beta,\theta,\tau^2)}{p(y)} \quad (7)$$
$$\propto p(\theta,\beta,\tau^2) \times p(y|\theta,\beta,\tau^2).$$

The prior distribution can be informative or noninformative. Noninformative priors typically deliver inference consistent with classical methods. Even for weakly informative priors, the inference is often close to classical methods because the effect of the data typically overwhelms the prior. While often producing inference numerically very similar to classical inference, Bayesian inference will retain simpler interpretability.

Suppose we have toxin measurements at points $\gamma(t_1)$, $\gamma(t_2)$, ..., $\gamma(t_n)$ on the coast and have collected the $y(t_i)$'s in an $n \times 1$ vector y. Let X be the $n \times p$ matrix with ith row $x^T(t_i)$ and ω be the $n \times 1$ vector with elements ω_i. The posterior distribution of the model parameters is

$$p(\beta,\sigma^2,\tau^2,\phi|y) \propto U(\phi|a_\phi,b_\phi) \times IG(\tau^2|a_{\tau^2},b_{\tau^2})$$
$$\times IG(\sigma^2|a_{\sigma^2},b_{\sigma^2}) \times N(\beta|\mu_\beta,V_\beta) \times N(y|X\beta+\omega,\tau^2 I), \quad (8)$$

where $U(\cdot,\cdot)$, $IG(\cdot,\cdot)$, and $N(\cdot,\cdot)$ represent the uniform, the inverse-Gamma and the Normal distributions, respectively, as expounded in Gelman *et al.* (2013).

Posterior distributions, in general, are not available in simple closed-forms. Instead we sample $\{\beta, \omega, \theta, \tau^2\}$ from their posterior distribution, where $\theta = \{\sigma^2, \phi\}$, using Markov chain Monte Carlo (MCMC) methods (Gelman *et al.*, 2013; Banerjee *et al.*, 2014). Some simplifications are often made. One is to use a flat completely noninformative prior on β. Another is to integrate out ω from equation (8). The posterior samples for $\{\beta, \sigma^2, \tau^2, \phi\}$ are then obtained by simulating

$$p(\beta,\theta,\tau^2|y) \propto U(\phi|a_\phi,b_\phi) \times IG(\tau^2|a_{\tau^2},b_{\tau^2})$$
$$\times IG(\sigma^2|a_{\sigma^2},b_{\sigma^2}) \times N(y|X\beta,K_\theta+\tau^2 I). \quad (9)$$

The posterior samples for ω are subsequently obtained by sampling one instance of ω from $N(\cdot,\cdot)$ for each

sampled value of $\{\beta, \sigma^2, \tau^2, \phi\}$. This is called composition sampling (Banerjee *et al.*, 2014).

Suppose we have collected M post-convergence posterior samples for the model parameters, say $\{\beta_{(j)}, \theta_{(j)}, \tau_{(j)}^2\}$, for $j = 1, 2, ..., M$. The posterior samples for $Y(t_0)$ are obtained by composition sampling, i.e. for each j we draw $Y_{(j)}(t_0)$ from the conditional normal distribution, say $N(m_{(j)}, v_{(j)}^2)$, where the mean and variance are

$$m_{(j)}(t_0) = x(t_0)^T \beta_{(j)} + \tilde{K}_{\theta_{(j)}}(t_0,t) \tilde{K}_{\theta_{(j)}}^{-1}(t,t)$$
$$\tilde{K}_{\theta_{(j)}}(t,t_0)(y - X\beta_{(j)})$$
$$\text{and} \quad v_{(j)}^2(t_0) = \tilde{K}_{\theta_{(j)}}(t_0,t_0) - \tilde{K}_{\theta_{(j)}}$$
$$(t_0,t)\tilde{K}_{\theta_{(j)}}^{-1}(t,t)\tilde{K}_{\theta_{(j)}}(t,t_0), \quad (10)$$

where $\tilde{K}_{\theta_{(j)}}(\cdot,\cdot) = K_{\theta_{(j)}}(\cdot,\cdot) + \tau^2 I$. Note that $m_{(j)}(t_0)$ and $v_{(j)}^2(t_0)$ are precisely the classical kriging estimator and variance evaluated at $\{\beta_{(j)}, \theta_{(j)}, \tau_{(j)}^2\}$. Bayesian kriging, therefore, quantifies uncertainty in kriging by averaging the classical kriging estimator over the posterior distribution of the parameters. The resulting $Y_{(j)}(t_0)$ are samples from the posterior predictive distribution. The mean of these samples yields a point estimate of the predicted value at t_0, while the variance of the posterior samples estimates the predictive variance.

One assumption to simplify matters is that ϕ and $\alpha = \dfrac{\tau^2}{\sigma^2}$ are fixed, say at values resulting from the empirical variogram (Banerjee *et al.*, 2014). Hence, the posterior samples for the model parameters are obtained from the conjugate model

$$p(\beta,\sigma^2|y) \propto IG(\sigma^2|a_{\sigma^2},b_{\sigma^2})$$
$$\times N(\beta|\mu_\beta,\sigma^2 V_\beta) \times N(y|X\beta,\sigma^2 V_y), \quad (11)$$

where $V_y = R(\phi) + \alpha I$ and $R(\phi)$ is the spatial correlation matrix with elements $\exp(-\phi|t_i - t_j|)$. Here one can sample exactly from the posterior distribution in equation (11). For each $j = 1, 2, ..., M$, we first draw $\sigma_{(j)}^2 \sim IG(a_{(j)}^*, b_{(j)}^*)$ followed by $\beta_{(j)}|\sigma_{(j)}^2, y \sim N(Bb, B\sigma_{(j)}^2)$, where $a_{(j)}^* = a_{\sigma^2} + n/2$ and $b^* = b_{\sigma^2} + (y^T V_y y - b^T B b)/2$, where $B = (X^T V_y^{-1} X + V_\beta^{-1})^{-1}$ and $b = X^T V_y^{-1} y$.

Simulation

The simulated data consists of $n = 100$ data points. The outcome $Y(t)$ values were generated on an ellipse. We first generated $l_i \sim Uni(0, 2\pi)$ for $i = 1, 2, ..., n$, where the corresponding parametric equations are $m = 2\cos(l)$ and $n = \sin(l)$. We then drew a multivariate normal random

variable $\omega \sim N(0, K_\theta)$ and then $y(t_i) \sim N(\beta_0 + \omega(t_i), \tau^2)$, where t_i is the arc-length between points (m_{i-1}, n_{i-1}) and (m_i, n_i).

In the data generation step, we fixed $\tau^2 = 0.1$, $\beta = 0$, and $\theta = \{1, 1\}$. For assessing predictive performance, we used 75 observations for training the model and withheld 25 observations for testing the predictive validation.

We estimated the models in equations (9) and (11). To compare the performance of coastal kriging to kriging using Euclidean distance, we estimated the model in equation (1) as well using the covariance in equation (2). For all models, we assigned a noninformative prior to β_0 (i.e. $V_\beta^{-1} = O$ the matrix of zeroes) and an $IG(2, 2)$ prior to τ^2. In equation (9), σ^2 and ϕ were assigned $IG(2,2)$ and $U(0.8, 30)$ priors. The $IG(2,b)$ prior provides a prior mean of b but has, in theory, an infinite variance yielding a relatively vague prior but with a prior value centered around b. In equation (11), we fixed $\phi = 1.07$ and $\alpha = 0.25$ for the coastal kriging model and $\phi = 22009.68$ and $\alpha = 8.13 \times 10^{-5}$ for kriging with Euclidean distance. Starting values for σ^2, τ^2, and ϕ in equation (8) and the fixed values for ϕ and $\alpha = \tau^2/\sigma^2$ in equation (11) were provided using their estimates from the empirical variogram for the data (Banerjee *et al.*, 2014).

We also compared coastal kriging to universal kriging (UK). UK is kriging with a trend, where $E(Z(s))$ is a linear combination of the known functions $\{f_0(s), \ldots, f_p(s)\}$ (Cressie, 1993). We assume that the mean $E(Z(s))$ is a function of the coordinates in a linear form, i.e. $Z(s) = \beta_0 + \beta_1 x_1(s) + \beta_2 x_2(s) + \omega(s) + \varepsilon(s)$, where $x_1(s)$ is the longitude at location s, $x_2(s)$ is the latitude at location s. In practice, we will not have an exact parametric formula for the coastline. This needs to be approximated by simple parametric curves. The easiest such option is a sequence of line segments, as described earlier. We used our simulated dataset to evaluate the performance of such linear approximations.

Let $\Delta m_i = m_i - m_{i-1}$ and $\Delta n_i = n_i - n_{i-1}$, then the length of the straight line segment connecting the two points is $t^* = \sqrt{(\Delta m)^2 + (\Delta n)^2}$. For small Δm, the sum of the lengths of these line segments provides an approximation to the length of the curve. We will, therefore, consider four models for coastal kriging. The model in equation (9) with the exact parametrization for an ellipse will be called Model 1a, while that with linear approximation will be called Model 1b. Similarly, the exact and approximate parameterizations corresponding to the model in equation (11) will be referred to as Model 2a and Model 2b, respectively.

Table 1 presents the posterior medians and 95% Bayesian credible intervals for the parameters in each of the above four models, the simple Euclidean distance kriging model and the UK model. The credible intervals from all models include the true values of β_0. Models

Table 1. Medians, 2.5% and 97.5% quantiles of the posterior samples of the coefficient estimate, partial sill σ^2, nugget effect τ^2, decay parameter ϕ, MSPE, DIC, K–L, and CV(10) for the fitted models to the simulated data.

	True	Model 1a[a]	Model 1b[b]	Model 2a[c]	Model 2b[d]	Simple kriging[e]	UK
β_0	0	0.16 (−0.41, 0.79)	0.16 (−0.51, 0.89)	0.19 (−0.35, 0.71)	0.24 (−0.49, 0.97)	0.006 (−0.16, 0.18)	0.03 (−0.24, 0.31)
σ^2	1	0.55 (0.30, 1.09)	0.56 (0.30, 1.14)	0.48 (0.36, 0.67)	0.74 (0.55, 1.04)	0.58 (0.43, 0.81)	0.18 (0.12, 0.26)
τ^2	0.1	0.18 (0.12, 0.28)	0.18 (0.12, 0.27)	0.12 (0.09, 0.17)	0.12 (0.09, 0.17)	2.8×10^{-5} (2.1×10^{-5}, 3.9×10^{-5})	0.17 (0.12, 0.25)
ϕ	1	1.20 (0.85, 2.81)	1.15 (0.71, 3.58)	0.76	0.76	31773.42	0.32 (0.16, 0.96)
MSPE		0.57	0.59	0.53	0.54	1.23	0.56
DIC		30.11	30.95	28.80	29.82	55.43	37.8
K–L		4.10	5.61	5.68	5.64	73.67	100.7
CV(10)		0.170	0.169	0.171	0.176	0.558	0.183

[a]Full hierarchical model using arc-length.
[b]Full hierarchical model using line segment approximation.
[c]Simplified hierarchical model using arc-length.
[d]Simplified hierarchical model using line segment approximation.
[e]Simplified hierarchical model using Euclidean distance.

1a and 1b captured the true values of σ^2 and ϕ. Model 2b also captured the true value of σ^2 and Models 2a and 2b captured the true value of τ^2. To assess predictive performance across the six models, we used mean square prediction error (MSPE). Coastal kriging and UK models produced very similar MSPE values, and the highest MSPE was produced by the simple Euclidean distance kriging model. For model comparison, we also used the Kullback–Leibler (K–L) divergence criterion $(D_{KL}(M_0|M_i))$, $i = 1,...5$, where M_0 is the true distribution and M_i is the distribution under model i. For multivariate normal distributions, the K–L divergence (Bedrick and Tsai, 1994) takes the form

$$\frac{1}{2}\left(\begin{array}{c} tr\left(\Sigma_{M_i}^{-1}\Sigma_{M_0}\right) + \left[X\beta_{M_i} - X\beta_{M_0}\right]^T \Sigma_{M_i}^{-1}\left[X\beta_{M_i} - X\beta_{M_0}\right] \\ -n + \ln\left(\det\left(\Sigma_{M_i}\right)\right) - \ln\left(\det\left(\Sigma_{M_0}\right)\right) \end{array} \right),$$

where $\Sigma = K_\theta + \tau^2 I$. Model 1a produced the lowest D_{KL} followed by Models 1b, 2b, and 2a, and the highest values were produced by the UK model and the simple Euclidean distance kriging model. We also used deviance information criterion (DIC), which is commonly used in Bayesian model selection. Model 2a produced the lowest value, followed by Models 2b, 1a, and 1b, and the UK model. The highest (worst performing) value was produced by the simple Euclidean distance kriging model. Finally, 10-fold cross-validation [CV(10)] was the lowest among coastal kriging models followed by the UK model then the simple Euclidean distance kriging model.

Figure 1 presents the 'coastal correlation', i.e. the correlation function plotted against the distance along the coast. The solid line represents the posterior mean of

$K_\theta(s, s')/\sigma^2$, while the shaded region represents the corresponding 95% credible band providing uncertainty quantification for the spatial covariance using Model 1a. We also used Bayesian 95% prediction intervals and the predicted mean values of the outcome from the 25 holdout locations and plotted them against the true values; the results are seen in Fig. 2. For the coastal kriging models, the intervals include the true values of the outcome variable in each of the holdout locations except for one location. The UK model provided improved prediction over simple Euclidean distance kriging model which produced the least accurate prediction with wider credible intervals.

These results indicate that Bayesian models using piecewise linear approximations to a parametric curve do not seem to adversely affect the inferential performance relative to models using the true form of the parametric curve. They also indicate that coastal kriging is better than classical kriging methods such as simple Euclidean distance kriging and UK when the source of variability in the data arises from a curve. Thus, Models 1b and 2b are good candidate models to be used in the data analysis (Fig. 3).

Data analysis

Coastal kriging of the concentration of chemicals inhaled by the clean-up workers following the oil spill in 2010 may be useful to assess the potential health effects associated with the spill for locations without measurements. The data set used here consists of air samples collected on clean-up workers on Waveland Beach, Mississippi which extends in an S-shape for 7–8 km (Fig. 4). The samples were collected for ~10 h per day using passive
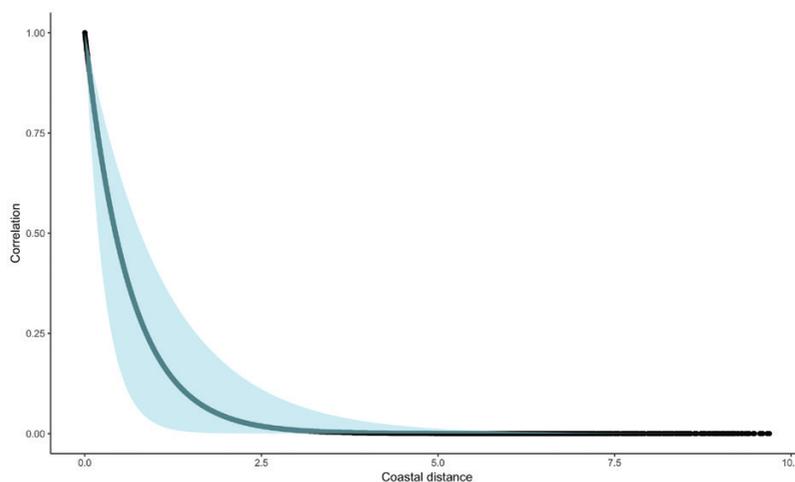

Figure 1. Coastal kriging estimated correlation (solid line) versus coastal distance applying Model 1a to the simulated data along with 95% credible bands (shaded).
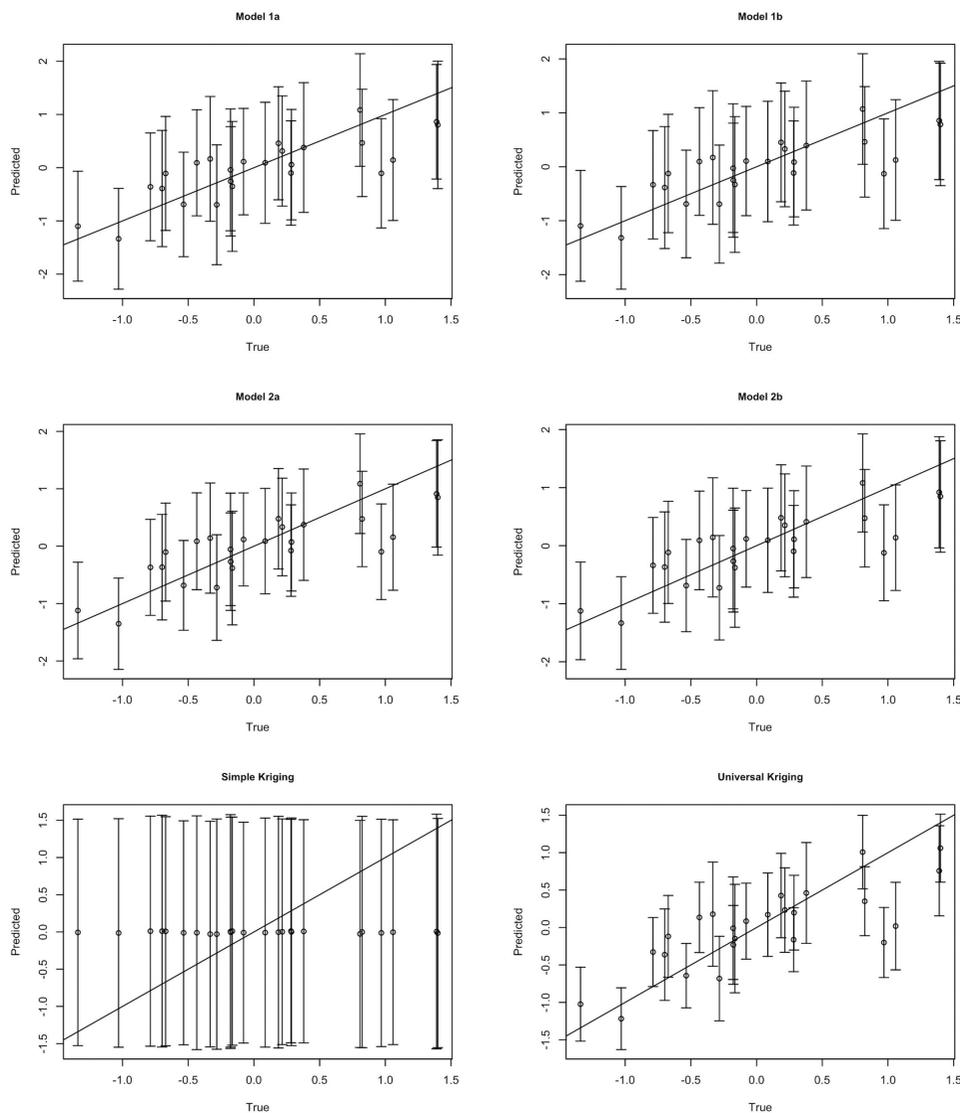
Figure 2. Simulated data true versus predicted values with 95% prediction intervals with 45° line of the four coastal kriging and simple kriging models.

dosimeters clipped to the workers' collars to measure breathing zone concentrations. The chemicals in the air diffused on to a charcoal pad inside the sampler. Five analytes were analyzed at the laboratory. They include total hydrocarbons (THC) which is a composite of the volatile chemicals in crude oil and is our main variable of interest. There were a total of 60 sample points [THC parts per million (ppm)] collected between 19 September and 21 December 2010 that were used in the analysis. Two exposure groups were considered, workers who cleaned jetties and other land-based structures and workers who cleaned beaches.

Candidate models include Models 1b and 2b where the curve is approximated by line segments and the parameterization in equation (3) is used. The fixed values of ϕ and α in equation (11) could be the estimated from the variogram. However, in coastal kriging, the variogram may not provide accurate estimates. Hence, we will use Model 1b in the data analysis and compare the results to simple Euclidean distance kriging results. For both models, we assigned a noninformative prior to β_0 (i.e. $V_\beta^{-1} = O$ the matrix of zeroes) and an $IG(2,2)$ prior to τ^2. In equation (9), σ^2 and ϕ were assigned $IG(2,2)$ and $U(0.8, 30)$ priors. The prior on ϕ implies
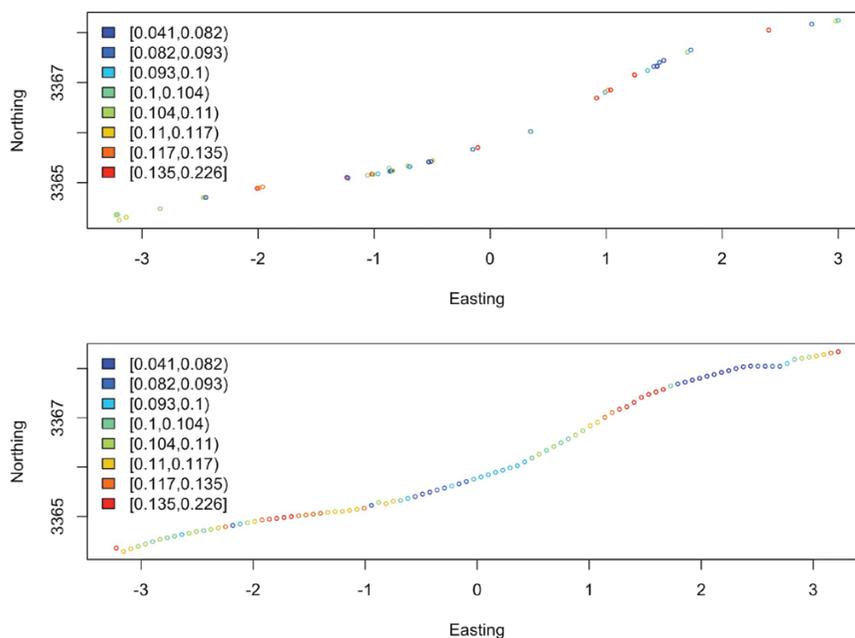
Figure 3. Observed THC (ppm) (top panel) and interpolated values from Model 1b (bottom panel) over Waveland Beach, Mississippi.

that the effective spatial range, i.e. the distance beyond which spatial correlation is negligible, is between 0.1 and 3.8 on a coastline with a distance of 7.6 km. In addition, coastal kriging was compared to UK with a linear trend.

Twelve observations acted as a holdout testing sample and the models were assessed based on their predictive performance at new locations using MSPE in addition to CV(10) and on the goodness of fit measure DIC. All observations were log transformed to achieve normality.

Table 2 shows parameter estimates of the fitted models. MSPE is almost the same among the three models, and the highest CV(10) resulted by the UK model. Model 1b produced the lowest DIC value. Results show that coastal kriging proposed in equation (5) provides a better fit for coastal data compared to other classical kriging methods. The top panel of 3 again shows the observed levels, while the bottom panel in 3 shows interpolated total hydrocarbon (ppm) values obtained from Model 1b (posterior predictive means) along a string of 100 coordinates on the coastline. These figures evince the effect of coastal kriging: the interpolated values are in close agreement with the observations at locations close to those with observations, while smoothing the values at unobserved points by borrowing from neighboring observations. Figure 4 shows a map of the observed *and*

interpolated measurements along the coastline overlaid on a GoogleMap with a legend indicating the level of the observed THC (ppm) over Waveland Beach, Mississippi.

Discussion

We developed a simple and flexible Bayesian framework for spatially oriented data that can be used to assess exposures of workers by interpolating levels of chemicals along a coastline. The statistical models for coastal kriging exploit a simple representation of the coast as a parametric function of the coordinates of points along the coastline. We presented four models using two different parameterizations. We found that for a simple curve, 'kriging' using line segment approximation performs better than spatial kriging using Euclidean distance. This could be a useful and practical approach for kriging over any simple curve. The model is relatively easy to fit since the covariance depends on parameters in $\Re^1$.

We remark that the current article only considers worker exposure assessment, not community-based exposure assessment. In the GuLF STUDY, >28 000 samples of THC and several other chemicals were collected across the Gulf, along the coasts, and at ports and docks, providing sufficient data for the STUDY exposure estimates (Stewart *et al.*, 2017). These estimates
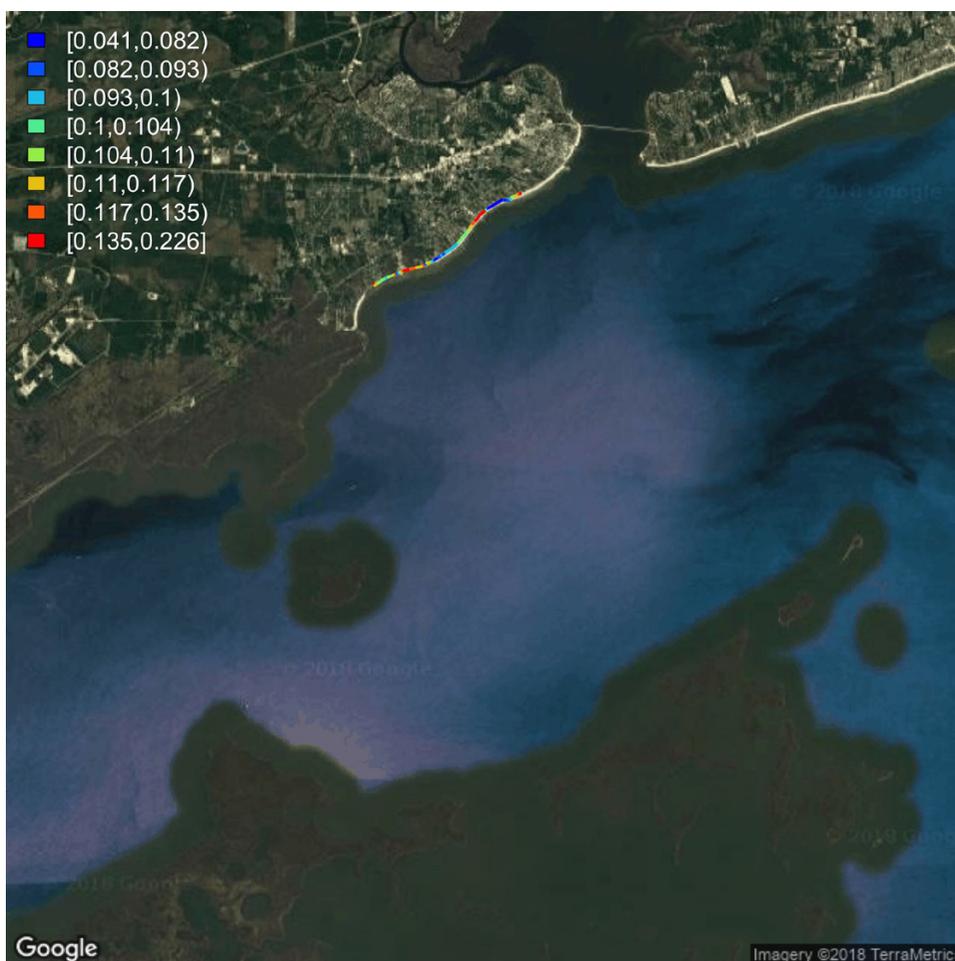
Figure 4. Map of observed and interpolated THC (ppm) from Model 1b over Waveland Beach, Mississippi.

Table 2. Medians, 2.5% and 97.5% quantiles of the posterior samples of the coefficient estimate, partial sill σ^2, nugget effect τ^2, decay parameter ϕ, MSPE, DIC, and CV(10) for the fitted models of the log-transformed THC.

	Model 1[a]	Simple kriging[b]	UK
β_0	−2.29 (−2.71, −1.83)	−2.23 (−2.67, −1.73)	−71.2 (−8663.9, 7497.1)
σ^2	0.59 (0.29, 1.26)	0.59 (0.28, 1.15)	0.52 (0.34, 0.89)
τ^2	0.46 (0.25, 0.80)	0.46 (0.27, 0.85)	0.17 (0.12, 0.23)
ϕ	9.08 (1.26, 24.82)	7.43 (1.78, 22.70)	0.29 (0.29, 6.48)
MSPE	0.06	0.06	0.05
DIC	34.4	38.6	65.05
CV(10)	0.06	0.06	0.13

[a]Full hierarchical model using line segment approximation.
[b]Full hierarchical model using Euclidean distance.

were derived from groups of samples based on the tasks being performed. The concentrations generated by these tasks (i.e. cleaning the beaches of oil and tar) represent task-derived exposures and, to a lesser extent, ambient air exposures. Using such task-based measurements is not appropriate to impute general or community air

concentrations because the task concentrations will be higher than ambient concentrations due to the workers being nearer to the source of the chemical emission than the community. With the data used here, however, the imputed concentrations from the methodology described above may represent workers' exposures performing those same tasks in unmeasured locations. To date, occupational assessment methodologies have focused primary on fairly localized exposure situations. The method described here may be useful in more geographically extended situations, such as workers building a highway or mitigating a chemical release in a river or residents living along a fenceline adjacent to a manufacturing site.

Our study has some limitations within which our findings need to be interpreted carefully. First, the results are based on a total of 60 data points from which 48 were used in training the model and 12 were used in testing it. Second, the data points are distributed on a coast with little curvature which rendered the coastal kriging results slightly better than simple Euclidean distance kriging results. Last, but not least, the distribution of THC in the air is unknown and its source is not arising from the coast which may add some uncertainty in the fitted model, although in our data this uncertainty is assumed to be minimal.

Building valid models for coastal kriging presents many new research opportunities. For instance, it would be of interest to develop a model for more complicated coastlines, perhaps along closed curves such as the coasts of an island. Future work will also consider the modeling and analysis of censored data, as is commonplace in exposure studies, due to measurements below the limits of detection. Also, our current computations were cheap due to the relatively small number of spatial locations. Spatial models become expensive to compute for larger datasets, as perhaps would be the case with the full GuLF STUDY databases. Here, more specialized high-dimensional Bayesian models can be exploited (e.g. Banerjee, 2017). Finally, we will also consider extending this work to exposure assessment for communities rather than individuals.

Supplementary Data

Supplementary data are available at *Annals of Work Exposures and Health* online.

Conflict of Interest

No conflicting interests are declared.

References

Banerjee S. (2005) On geodetic distance computations in spatial modeling. *Biometrics*; **61**: 617–25.

Banerjee S. (2017) High-dimensional Bayesian geostatistics. *Bayesian Anal*; **12**: 583–614.

Banerjee S, Carlin BP, Gelfand AE. (2014) *Hierarchical modeling and analysis for spatial data*. Boca Raton, FL: Chapman and Hall/CRC.

Bedrick EJ, Tsai C-L. (1994) Model selection for multivariate regression in small samples. *Biometrics*; **50**: 226–31.

Cressie NA. (1993) *Statistics for spatial data*. Revised edn. Hoboken, NJ: John Wiley & Sons, Inc.

Gelman A, Carlin JB, Stern HS *et al*. (2013) *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Hewett P, Logan P, Mulhausen J *et al*. (2006) Rating exposure control using Bayesian decision analysis. *J Occup Environ Hyg*; **3**: 568–81.

Hoef JMV, Peterson EE. (2010) A moving average approach for spatial statistical models of stream networks. *J Am Stat Assoc*; **105**: 6–18.

Kwok R, Engel L, Miller A *et al*.; the GuLF STUDY Research Team. (2018) The GuLF STUDY: a prospective study of persons involved in the Deepwater Horizon oil spill response and clean-up. *Environ Health Perspect*, **125**: 570–578. doi:10.1289/EHP715

Ramachandran G, Banerjee S, Vincent JH. (2003) Expert judgment and occupational hygiene. *Ann Occup Hyg*; **47**: 461–75.

Stewart J. (2012) *Multivariable calculus*. 7th edn. Pacific Grove, CA: Brooks/Cole.

Stewart P, Stenzel M, Ramachandran G *et al*. (2018) Development of a total hydrocarbon ordinal job-exposure matrix for workers responding to the Deepwater Horizon disaster: the GuLF STUDY. *J Expo Sci Environ Epidemiol*, **28**: 223–230. doi:10.1038/jes.2017.16.