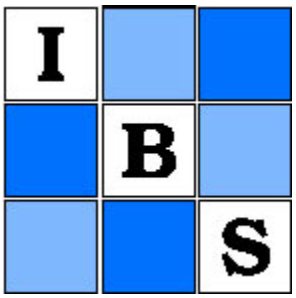

Bayesian Variogram Modeling for an Isotropic Spatial Process

Mark D. ECKER and Alan E. GELFAND

The variogram is a basic tool in geostatistics. In the case of an assumed isotropic process, it is used to compare variability of the difference between pairs of observations as a function of their distance. Customary approaches to variogram modeling create an empirical variogram and then fit a valid parametric or nonparametric variogram model to it.

Here we adopt a Bayesian approach to variogram modeling. In particular, we seek to analyze a recent dataset of scallop catches. We have the results of the analysis of an earlier dataset from the region to supply useful prior information. In addition, the Bayesian approach enables inference about any aspect of spatial dependence of interest rather than merely providing a fitted variogram. We utilize discrete mixtures of Bessel functions that allow a rich and flexible class of variogram models. To differentiate between models, we introduce a utility-based model choice criterion that encourages parsimony. We conclude with a fully Bayesian analysis of the scallop data.

Key Words: Bessel functions; Correlation functions; Importance sampling; Mixtures; Model determination; Stationary process

1. INTRODUCTION

Suppose we collect data that take the form of a partial realization of a random spatial process in R^n that is, we observe $Y(\mathbf{s}_i)$ at a set of locations $\mathbf{s}_i \in R^n$. It is assumed that $\text{var}(Y(\mathbf{s}_i) - Y(\mathbf{s}_j))$ is only a function of the separation vector $\mathbf{s}_i - \mathbf{s}_j$. If it depends only upon $\|\mathbf{s}_i - \mathbf{s}_j\|$, the distance between sites, the process is said to be *isotropic* and the function $2\gamma(\|\mathbf{s}_i - \mathbf{s}_j\|) = \text{var}(Y(\mathbf{s}_i) - Y(\mathbf{s}_j))$ is called the *variogram*. Typically, the variogram is assumed to increase in $\|\mathbf{s}_i - \mathbf{s}_j\|$, the rationale being that the differences between pairs of observations closer in space should tend to exhibit less variability than that for pairs farther apart.

The variogram only determines one feature of the spatial process $Y(\mathbf{s})$. To implement likelihood or Bayesian inference requires specification of the joint distribution of an arbitrary set $Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_N)$. This is typically done by assuming $Y(\mathbf{s})$ to come from a stationary Gaussian process with constant mean and isotropic covariance function

---

Mark D. Ecker is Assistant Professor, Department of Mathematics, University of Northern Iowa, Cedar Falls, IA 50614-0506. Alan E. Gelfand is Professor, Department of Statistics, University of Connecticut, Storrs, CT 06269-3120.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 2, Number 4, Pages 347–369

$\text{cov}(Y(\mathbf{s}_i), Y(\mathbf{s}_j)) = C(\|\mathbf{s}_i - \mathbf{s}_j\|)$. Then C determines the variogram 2γ, that is, $2\gamma(\|\mathbf{s}_i - \mathbf{s}_j\|) = 2(C(0) - C(\|\mathbf{s}_i - \mathbf{s}_j\|))$. In general, γ does not determine C because stationary increments for $Y(\mathbf{s}_i) - Y(\mathbf{s}_j)$ (the so-called *intrinsic hypothesis*) do not imply that $Y(\mathbf{s})$ is stationary. In fact, $C(0)$ need not exist.

Given a parametric model for γ or C, techniques such as maximum likelihood (Cressie 1993, sec. 2.6.1), weighted least squares (Cressie 1985), or fitting by inspection are usually used to estimate the model parameters. We adopt a Bayesian approach for several reasons. First, for the dataset of interest, it is sensible to use information gained through a previously studied dataset from the region. No current variogram modeling technique routinely incorporates prior information. Second, the Bayesian paradigm allows finite sample inference regarding the variability of the variogram parameters. Maximum likelihood techniques rely on asymptotic results. Weighted least squares estimates are dependent upon the arbitrary construction of the variogram (see Cressie 1993, sec. 2.6.2). Fitting by inspection is not sound statistical practice and provides no estimation of variability. Third, the Bayesian paradigm provides an entire posterior distribution for each variogram parameter, avoiding possibly inappropriate approximate normality assumptions. In addition, a fully Bayesian analysis allows inference (again an entire posterior distribution) for any aspects of the variogram of interest (e.g., the sill, the nugget, and the range). It also allows a confidence band for the variogram itself. In return for a full parametric distributional specification for the spatial process $Y(\mathbf{s})$, full inference is available. If only a parametric form for γ is specified and then fit to a particular empirical variogram, inference is limited to a point estimate of γ.

Our work is motivated by two datasets consisting of scallop counts off the New Jersey and Long Island coastline. Since 1982, the Northeast Fisheries Center of the National Marine Fisheries Service in Woods Hole, Massachusetts, has annually sampled on the continental shelf off the Northeastern United States to estimate the abundance of sea scallops and other shellfish. Their methodology is to stratify the region from the Georges Bank to Cape Hatteras based upon water depth and latitude. One geographical region of interest is the New York Bight, which encompasses an area in the Atlantic Ocean from the mouth of the Delaware River to the eastern tip of Long Island.

A total of 148 sites, sampled in 1990 in the New York Bight, were previously analyzed in Ecker and Heltshe (1994) and are readily available as part of the S-Plus SpatialStats version 1.0 module. Subsequently in 1993, 147 different locations were sampled in this region. Figure 1 displays the sampled locations for both 1990 and 1993, while Figure 2 provides a gray-scale display for the log-scaled scallop counts for the 1993 data. The darker the shading, the higher the counts.

The class of all permissible variogram models in R^n has been characterized (see Schoenberg 1938, theorem 1). In R^2, the allowable class for isotropic covariance structures is mixtures of Bessel functions of the first kind of order zero. Thus, we propose finite mixtures of such functions to provide an arbitrarily rich covariance specification, hence, a very flexible class of variogram models. Similar strategies in the literature include Sampson and Guttorp (1992), who used mixtures of Gaussian type correlations, and Shapiro and Botha (1991) and Cherry, Banfield, and Quimby (1996), who considered discrete mixtures of Bessel functions when $n = 2$. However, all of this work uses least squares fitting of these mixture models to an arbitrarily constructed empirical variogram.

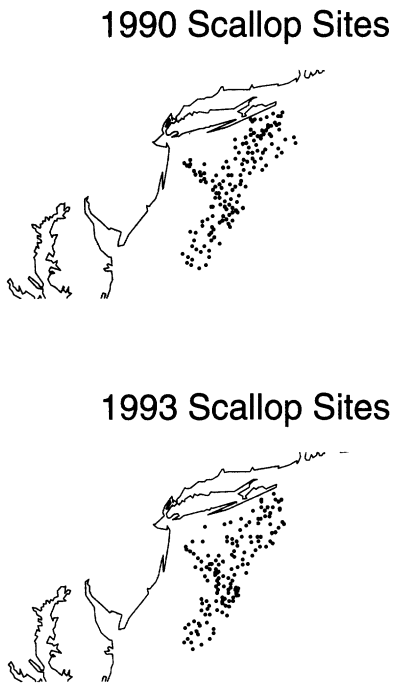

Figure 1. Sites Sampled in the Atlantic Ocean for 1990 and 1993 Scallop Catch Data.

Instead, we insert these forms into the likelihood, enabling inference conditional on all the observed data, not on a particular summary.

One aspect of variogram modeling infrequently addressed in the literature is which of a set of fitted models best explains the data. Often it is argued that, based upon practical experience, a particular variogram specification is appropriate for a particular form of spatial data. Although empirical feel is invaluable, it would be useful to provide firm support, in terms of suitable predictive performance, for choosing one variogram model over another. A classical approach uses a goodness-of-fit criteria perhaps with a penalty for model dimension, as in the Akaike Information Criterion (AIC) (1973) or the Schwartz or Bayesian information criterion (BIC) (1978). See Webster and McBratney (1989) for an adaptation of the AIC and BIC methodologies to variogram modeling. Under the Bayesian perspective, we adopt the utility-based approach of Gelfand and Ghosh (in press), which minimizes a so-called balanced expected loss with respect to the posterior predictive distribution. Such a loss penalizes actions that depart from what we have observed, but also from what we expect to observe under the model. In the context of variogram estimation, a related point arises. The signal (the true variogram) is often accompanied by enormous noise (variability about the true variogram), making virtually any plausible variogram model fit poorly. We devise a method to sharpen model choice given this difficulty.

In Section 2, we review standard variogram models and describe the relationship between the variogram and the correlation structure of the process. In Section 3, we examine the class of all valid variogram models in R^2. The computationally intensive

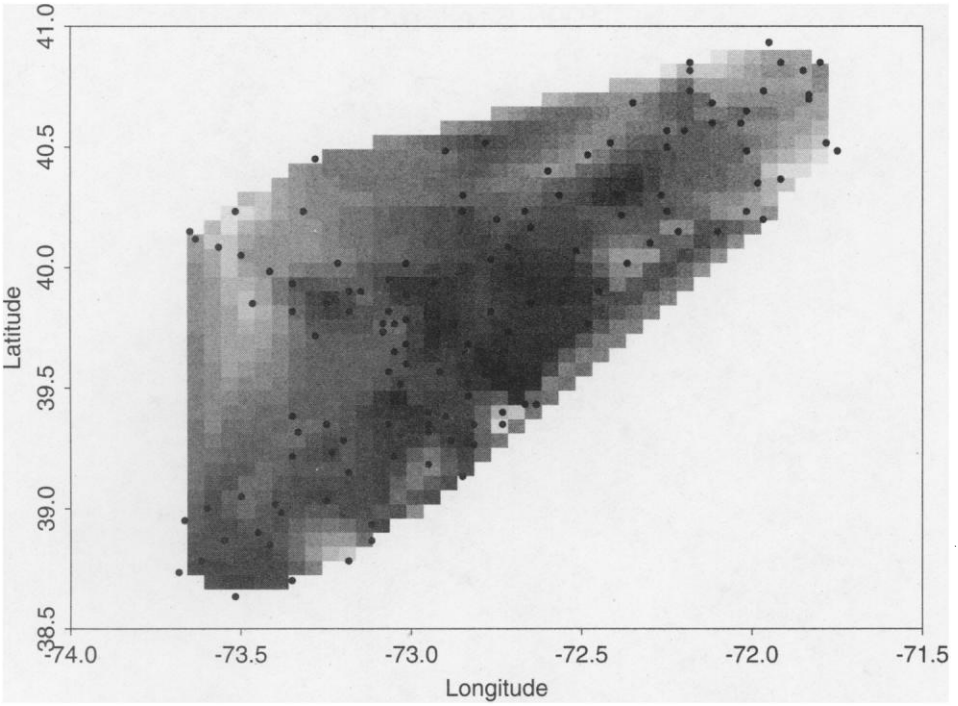

Figure 2. Smoothed Log-Scaled Scallop Counts in 1993. The darker the shading the larger the counts.

Bayesian model-fitting procedure, outlined in Section 4, is used to estimate the model parameters. In Section 5, the aforementioned model choice criterion is developed. We also clarify the signal-to-noise problem and suggest a remedy. Section 6 analyzes the scallop data mentioned, and Section 7 provides conclusions.

2. ISOTROPIC VARIOGRAM MODELS

For sites $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$ in R^n, let $\mathbf{y} = (Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_N))'$ be the response vector. Under the intrinsic hypothesis of Matheron (1963), the increments $Y(\mathbf{s}_i) - Y(\mathbf{s}_j)$ are stationary, that is,

$$\begin{aligned} \mathrm{E}(Y(\mathbf{s}_i) - Y(\mathbf{s}_j)) &= 0 \\ \mathrm{var}(Y(\mathbf{s}_i) - Y(\mathbf{s}_j)) &= 2\gamma(\mathbf{s}_i - \mathbf{s}_j). \end{aligned} \tag{2.1}$$

When γ is invariant to rotations of the spatial coordinates (i.e., the spatial variability is the same for any direction), the process is said to be isotropic, in which case γ depends only upon d_{ij}, the Euclidean distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$. $2\gamma(d_{ij})$ is called the variogram of the process while $\gamma(d_{ij})$ is the *semivariogram*. If we further assume that $\mathbf{y}$ is second-order stationary, then $\mathrm{E}(Y(\mathbf{s}_i)) = \mu$ and $\mathrm{cov}(Y(\mathbf{s}_i), Y(\mathbf{s}_j)) = C(d_{ij}) < \infty$ is also a function of the Euclidean distance between sites.

The fitting of models for $\gamma(d_{ij})$ takes two routes. One is a nonparametric methodology that imposes no distribution on $\mathbf{y}$ and views variogram modeling as a curve-

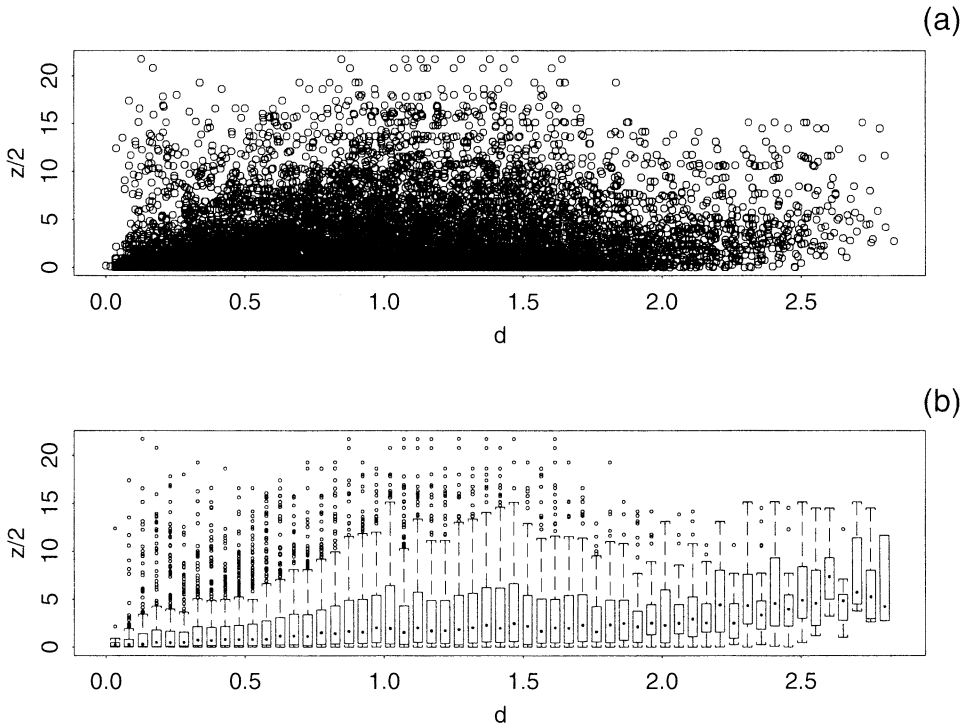

Figure 3. Semivariogram Cloud (a) for 1993 Scallop Data, and Boxplot Produced From .05 Lag (b).

fitting exercise. For each of the $N(N-1)/2$ pairs of data points in R^n, a plot of $[(Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2]/2$ versus d_{ij} is termed the *semivariogram cloud*. The semivariogram cloud for the 1993 scallop data apears in Figure 3(a). The Matheron (1963) estimator of the semivariogram,

$$\gamma_r^*(d_{ij}) = \frac{1}{2N_{B_r}} \sum_{\{(i,j):d_{ij}\in B_r\}} (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2, \tag{2.2}$$

smooths the semivariogram cloud by aggregating distances into R sets $B_1, B_2, \ldots, B_R$. $B_r = \{d_{ij} : b_{r-1} \le d_{ij} < b_r\}$ for constants b_{r-1} and $b_r, r = 1, \ldots, R$ with $b_0 = 0$, and N_{B_r} equals the number of d_{ij} in B_r with $\sum_r N_{B_r} = N(N-1)/2$. Usually, the constants

Table 1. Common Parametric Correlation Forms

Name	$\rho\ (d, \phi)$	
Exponential	$\exp(-\phi d)$	
Gaussian	$\exp(-\phi\ d^2)$	
Cauchy	$1/(1+\phi d^2)$	
Bessel	$J_0(\phi\ d)$	
Spherical	$\frac{1}{2}(\phi^3\ d^3 - 3\phi\ d + 2)$	if $d \le \frac{1}{\phi}$
	0	if $d > \frac{1}{\phi}$

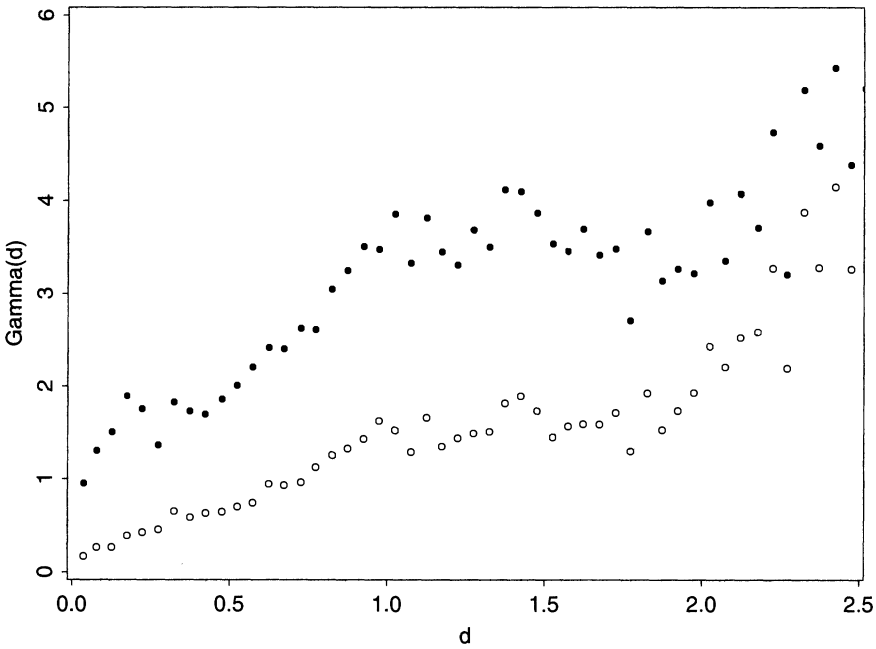

Figure 4. Matheron and Deviance Empirical Semivariograms for Lag δ = .05. Matheron = •; deviance = ○.

b_r are chosen to be $r\delta$ where δ is a specified *lag*. A plot of γ_r^* versus the midpoint of the interval (b_{r-1}, b_r) for $r = 1, \ldots, R$ is called the Matheron *empirical semivariogram*. It need not be a *valid* semivariogram; that is, γ^* need not be conditionally nonnegative definite [e.g., see Armstrong and Diamond (1984) and Christakos (1984)]. The Matheron empirical semivariogram for the 1993 scallop data appears in Figure 4.

The second route adopts a parametric specification for $\mathbf{y}$, customarily assuming that it is second-order stationary and normally distributed. Then we can write as in, for example, Diggle, Liang, and Zeger (1994, p. 87),

$$\mathbf{y} \sim \mathrm{N}(\mu\mathbf{1}, \Sigma(\boldsymbol{\alpha})), \tag{2.3}$$

where $\boldsymbol{\alpha} = [\tau^2 \ \sigma^2 \ \boldsymbol{\phi}]'$ and $\Sigma(\boldsymbol{\alpha}) = \tau^2 I + \sigma^2 H(\boldsymbol{\phi})$, with $(H(\boldsymbol{\phi}))_{ij} = \rho(d_{ij}, \boldsymbol{\phi})$ a valid parametric correlation function depending upon the distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$. Examples of standard parametric forms of $\rho(d, \boldsymbol{\phi})$ are given in Table 1, where ϕ becomes a scalar capturing the rate of correlation decay. $\Sigma(\boldsymbol{\alpha})$ determines the variogram model, and likelihood-based techniques can be used to estimate model parameters and to compare models.

For the scallop data described in the introduction, we take the response $Y(\mathbf{s}_i)$ to be log(total catch at $\mathbf{s}_i + 1$) where the constant 1 is added to address the observed zero catches. (We assume that zero catches do not arise because the site is unsuitable for scallop habitation. Under the latter asumption a different modeling approach incorporating a point mass at zero might be appropriate.) Both Ecker and Heltshe (1994) and Kaluzny, Vega, Cardosa, and Shelly (1996) modeled the 1990 log transformed scallop catches as approximately normally distributed with justification by exploratory data analysis.

With nongeographical covariates, one can generalize the mean structure in (2.3) to a linear form, $\mathrm{E}(Y(\mathbf{s}_i)) = \sum_{j=0}^{p} X(\mathbf{s}_i)_j \beta_j$ with $X(\mathbf{s}_i)_0 = 1$. With only geographical covariates (as in the case of the scallop data), the mean is often modeled as a *trend surface*, a polynomial in the spatial locations. However, tension arises when capturing spatial effects through both the mean and the variance structure (e.g., see Journel and Rossi 1989), so here we confine ourselves to a constant mean.

The variogram resulting from (2.3) is given by

$$2\gamma(d, \boldsymbol{\alpha}) = 2\gamma(d, \tau^2, \sigma^2, \phi) = 2\left(\tau^2 + \sigma^2\left(1 - \rho(d, \phi)\right)\right). \tag{2.4}$$

The parameters τ^2, σ^2, and ϕ have a geostatistical interpretation. If $\rho(d, \phi) \to 0$ as $d \to \infty$, then the *sill* of the variogram is defined to be $\lim_{d\to\infty} 2\gamma(d) = 2(\tau^2 + \sigma^2)$ and represents the stationary variance of the process. Parametric forms such as the spherical achieve their sill for finite d, while the exponential, Gaussian, and Cauchy (rational quadratic) reach their sills asymptotically. For monotonically increasing variograms that reach their sill exactly, the *range* is defined to be the distance at which the process reaches its sill or, equivalently, the distance at which $\rho(d, \phi)$ becomes 0. Intuitively, points separated by distances greater than the range are spatially uncorrelated. For asymptotically silled variograms, two points will only be spatially uncorrelated in the limit as $d \to \infty$. Here, we speak of the *effective range* using two possible definitions. The first (McBratney and Webster 1986, p. 623) defines the range, r_C, as the distance d where the variogram reaches $2(\tau^2 + 0.95\sigma^2)$; that is, the spatial correlation of the process is 5%; and r_C solves $\rho(r_C, \phi) = .05$. The second (Cressie 1993, pp. 67–68) defines the range, r_V, to be the distance where the variogram reaches 95% of its sill; r_V solves $\rho(r_V, \phi) = .05(\tau^2 + \sigma^2)/\sigma^2$. For nonmonotonic variograms such as the Bessel, the range is not defined. The *nugget* of the variogram is $\lim_{d\to 0} 2\gamma(d) = 2\tau^2$, which, because of measurement error or a microscale effect resulting from extrapolating the variogram from the minimum sampled distance to the origin, need not be 0.

3. EXTENSIONS OF PARAMETRIC VARIOGRAM FORMS

A correlation function $\rho(d, \phi)$ is permissible (valid) only if it is positive definite in d, $\rho(0, \phi) = 1$, and $|\rho(d, \phi)| \leq 1, \forall d$. From Bochner's theorem (Cressie 1993, p. 84), the characteristic function of a symmetric distribution in R^n satisfies these constraints. For example, because $\exp(-\phi d^2)$ is the characteristic function of a $\mathrm{N}_n(0, 2\phi I)$ random variable, the Gaussian variogram is valid in any dimension. Feller (1966, p. 476) showed that the exponential and Cauchy, among others, are permissible in R^1. For validity of the spherical and Bessel, see Montoglou and Wilson (1982, p. 1381). The closure property of characteristic functions (Feller 1966, p. 477) states that a convex combination of characteristic functions is itself a characteristic function, extending $\rho(d, \phi)$ to mixture forms. The Pólya criterion (Chung 1974, pp. 182-183) provides sufficient conditions for recognizing permissible forms of $\rho(d, \phi)$ in R^1. Christakos (1984) and Armstrong and Diamond (1984) provided rigorous criteria for testing the validity of an arbitrary form $\rho(d, \phi)$ in R^n. If $\rho(d, \phi)$ is valid in R^{n_1}, then for any $n_2 < n_1$, $\rho(d, \phi)$ is valid in R^{n_2}. The converse in not necessarily true. See Armstrong and Jabin (1981) for a simple

counterexample.

From Khinchin's Theorem (Yaglom 1987, p. 106), the class of all valid functions $\rho(d, \phi)$ in R^n can be expressed (see also Schoenberg 1938) as

$$\rho(d, \phi) = \int_0^\infty \Omega_n(zd)\, dG_\phi(z), \tag{3.1}$$

where G_ϕ is nondecreasing integrable and

$$\Omega_n(x) = \left(\frac{2}{x}\right)^{\frac{n-2}{2}} \Gamma\left(\frac{n}{2}\right) J_{(\frac{n-2}{2})}(x).$$

Here, $J_v(\cdot)$ is the Bessel function of the first kind of order v. For $n = 1$, $\Omega_1(x) = \cos(x)$; for $n = 2$, $\Omega_2(x) = J_0(x)$; for $n = 3$, $\Omega_3(x) = \sin(x)/x$; for $n = 4$, $\Omega_4(x) = (2/x)J_1(x)$; and for $n = \infty$, $\Omega_\infty(x) = \exp(-x^2)$. Specifically,

$$J_0(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!^2} \left(\frac{x}{2}\right)^{2k},$$

and $\rho(d, \phi) = \int_0^\infty J_0(zd)\, dG_\phi(z)$ provides the class of all permissible variogram models in R^2.

In practice, a convenient, simple choice for $G_\phi(z)$ is a step function that assigns positive mass (jumps or weights) w_ℓ at points (nodes) ϕ_ℓ, $\ell = 1, \ldots, p$, yielding, with $\mathbf{w} = (w_1, w_2, \ldots, w_p)$,

$$\rho(d, \phi, \mathbf{w}) = \sum_{\ell=1}^{p} w_\ell \Omega_n(\phi_\ell d). \tag{3.2}$$

The forms in (3.2) are referred to as *nonparametric* variogram models in the literature, to distinguish them from standard or parametric forms for $\rho(d, \phi)$ such as those given in Table 1. This is a separate issue from selecting a parametric or nonparametric methodology for parameter estimation as discussed in Section 2. Sampson and Guttorp (SG) (1992), Shapiro and Botha (SB) (1991) and Cherry, Banfield, and Quimby (1996) used a step function for G_ϕ. Barry and Ver Hoef (1996) employed a mixture of piecewise linear variograms in R^1 and piecewise-planar models for sites in R^2. Hall, Fisher, and Hoffmann (1994) transformed the problem from choosing ϕ_ℓ's and w_ℓ's in (3.2) to determining a kernel function and its associated bandwidth. Lele (1995) proposed iterative spline smoothing of the variogram, yielding a ρ that is not obviously of the form (3.1).

Most of these nonparametric models are fitted to an arbitrarily lagged empirical variogram such as (2.2).

Sampson and Guttorp (1992) fitted their model, using $\Omega_\infty(x)$ in (3.2), to the semivariogram cloud rather than to the smoothed Matheron version given by (2.2). Their example involved a dataset with 12 sites, hence, yielding only 66 points in the semivariogram cloud, making this feasible. Application of their method to a much larger, hence, "noisier" dataset would be expected to produce a variogram mixing hundreds or perhaps thousands of Gaussian forms. The resulting variogram would follow the semivariogram cloud too closely to be plausible.

The SB method fixes the nodes ϕ_ℓ and estimates the jumps w_ℓ by an iterative weighted least squares fit of $2(1 - \rho(d, \boldsymbol{\phi}, \mathbf{w}))$ from (3.2) to the empirical variogram. Depending on the desired smoothness of the resulting variogram, this method offers three different sets of constraints. One set forces the estimated variogram to be concave; another imposes only monotonicity. The third forces only smoothness, from which the resulting variogram may again follow the empirical variogram values too closely. Cherry, Banfield, and Quimby (1996) evaluated the SB method and advocated $p = 200$ in (3.2).

Barry and Ver Hoef (BV) (1996) fitted linear-piecewise variograms to the empirical variogram and proved that, with enough components, their model could approximate any continuous variogram in R^1. Instead of mixing $\Omega_1(x) = \cos(x)$ curves in R^1, each defined over the entire support of the variogram as with the SG or SB methods, BV used individual linear components defined on disjoint subsets of the support. Hence, components can be focused on distance regions of higher interest, such as those near the origin.

Hall, Fisher, and Hoffman (1994) confined themselves to R^1. For the model in (3.1), assuming $G_{\boldsymbol{\phi}}(z)$ is differentiable, they let $dG_{\boldsymbol{\phi}}(z) = g(z)dz$ and employed a kernel estimator $\hat{g}$ of g. Their goal was to provide a valid variogram using $\Omega_1(x) = \cos(x)$, involving a continuous rather than discrete mixture. Their method thus requires choosing a kernel function for $\hat{g}$ and an associated bandwidth. Additionally, $\hat{g}$ needs to be "smoothed" to insure a positive-definite covariance matrix.

We work in R^2, where again $\Omega_2(x) = J_0(x)$. Within the Bayesian paradigm, we introduce (3.2) directly into the likelihood but keep p small (at most 5), allowing random w_ℓ or random ϕ_ℓ. We offer a compromise between the rather limiting standard parametric forms (Table 1) that specify three parameters for the covariance structure and the SB or SG method with the practically implausible mixture of hundreds of components. Moreover, by working with the likelihood, inference is conditioned upon the observed $\mathbf{y}$, rather than on a summary such as a smoothed version of the semivariogram cloud.

Returning to (3.1), when $n = 2$, we obtain

$$\rho(d, \boldsymbol{\phi}) = \int_0^\infty \sum_{k=0}^{\infty} \frac{(-1)^k}{k!^2} \left(\frac{zd}{2}\right)^{2k} dG_{\boldsymbol{\phi}}(z). \tag{3.3}$$

Only if z is bounded, that is, $G_{\boldsymbol{\phi}}$ places no mass on, say, $z > \phi_{\max}$, can we interchange summation and integration to obtain

$$\rho(d, \boldsymbol{\phi}) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!^2} \left(\frac{d}{2}\right)^{2k} \delta_{2k}, \tag{3.4}$$

where $\delta_{2k} = \int_0^{\phi_{\max}} z^{2k} dG_{\boldsymbol{\phi}}(z)$.

The simplest such choice for $G_{\boldsymbol{\phi}}$ puts discrete mass w_ℓ at a finite set of values $\phi_\ell \in (0, \phi_{\max}), \ell = 1, \ldots, p$, resulting in a finite mixture of Bessel's model for $\rho(d, \boldsymbol{\phi})$, which yields as in (2.4)

$$\gamma(d_{ij}) = \tau^2 + \sigma^2 \left(1 - \sum_{\ell=1}^{p} w_\ell J_0(\phi_\ell d_{ij})\right). \tag{3.5}$$

Under a Bayesian framework, for a given p, if the w_ℓ's are each fixed to be $1/p$, with ϕ_ℓ's unknown hence random, they are constrained by $0 < \phi_1 < \phi_2 < \ldots < \phi_p < \phi_{\max}$ for identifiability. The result is an equally weighted mixture of random curves. If a random mixture of fixed curves is desired, then the w_ℓ's are random and the ϕ_ℓ's are systematically chosen to be $\phi_\ell = [\ell/(p+1)]\phi_{\max}$. We examine $p = 2, 3, 4, 5$ for fixed nodes and $p = 1, 2, 3, 4, 5$ for fixed weights. Mixture models using random w_ℓ's and random ϕ_ℓ's might be considered but, in our limited experience, the posteriors have exhibited weak identifiability in the parameters and thus are not recommended.

In choosing $\phi_{\max}$, we essentially determine the maximum number of sign changes we allow for the dampened sinusoidal Bessel correlation function over the range of d's of interest. For say $0 \le d \le d_{ij}^{\max}$, the larger ϕ is the more sign changes $J_0(\phi d)$ will have over this range. This suggests making $\phi_{\max}$ very large. However, as noted earlier in this section, we seek to avoid practically implausible ρ and γ that would arise from an implausible $J_0(\phi d)$. For illustration, the plot in Figure 5 allows up to 11 sign changes (5 periods from the initial 0 of J_0). Letting κ be the value of x where $J_0(x) = 0$ attains its kth sign change [completes its $(k-1)/2$ period], we set $\kappa = \phi_{\max} d_{ij}^{\max}$, thus determining $\phi_{\max}$. We reduce the choice of $\phi_{\max}$ to choosing the maximum number of Bessel periods allowable. For a given p, when the ϕ's are random, the posterior distribution for ϕ_p will reveal how close to $\phi_{\max}$ the data encourages ϕ_p to be.

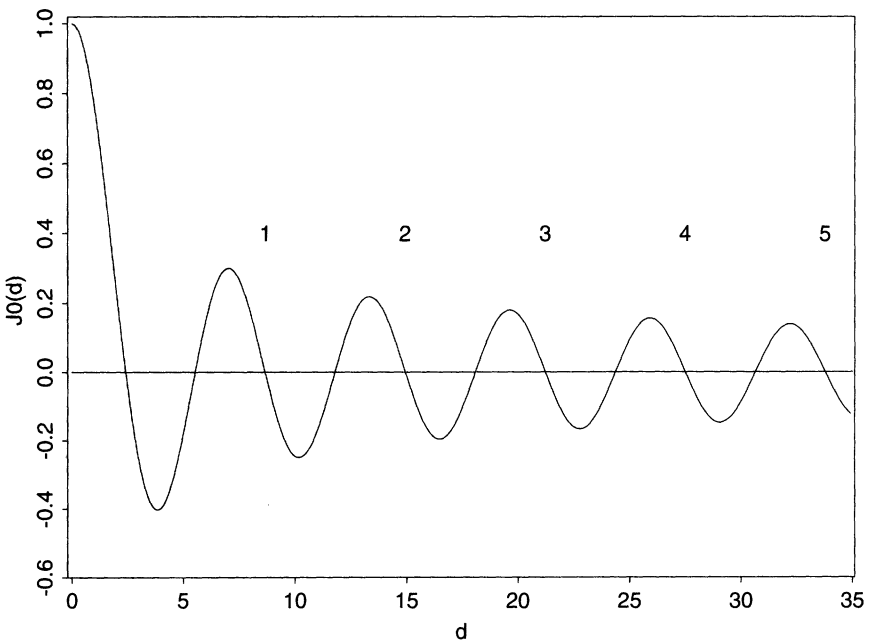

Figure 5. A Plot of J_0 Out to 11 Sign Changes (5 periods).

4. BAYESIAN MODEL FITTING

A handfull of recent papers have dealt with modeling spatial data from a Bayesian perspective. These include Handcock and Stein (1993) and Handcock and Wallis (1994), who modeled with the Matern class of correlation functions; Gaudard, Karson, Linder, and Sinha (1995), who used a mixture of exponential and Gaussian forms; and DeOliveira, Kedem, and Short (1997), who employed the general exponential correlation function. All focused on prediction, while our emphasis is on explanation of the spatial correlation structure. Although the two issues are often related, the model that best fits the data (our focus) need not be the model that best predicts responses at unsampled sites.

Denoting $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha})$ as the set of all model parameters, the likelihood $L(\boldsymbol{\theta}; \mathbf{y})$ is obtained through (2.3) using (3.2). The Bayesian model specification requires prior distributions for μ and $\boldsymbol{\alpha}$. For the parametric models of Table 1, we assume that the prior $\pi(\mu, \boldsymbol{\alpha})$ takes the form

$$\pi(\mu, \boldsymbol{\alpha}) = \pi_1(\mu)\pi_2(\tau^2)\pi_3(\sigma^2)\pi_4(\phi).$$

Although the parameters μ, τ^2, σ^2, and ϕ are not truly thought to be independent, the alternative, specifying a joint prior incorporating dependence, is arbitrary and difficult to justify. We prefer to let the data modify our independence assumption through the posterior. In fact, we prefer to let the data drive our inference, so for the covariance parameters, we assume rather uninformative inverse gamma (IG) distributions by setting the shape parameter equal to 2 (implying an infinite variance), that is, $\tau^2 \sim \text{IG}(2, b_{\tau^2})$, $\sigma^2 \sim \text{IG}(2, b_{\sigma^2})$ and $\phi \sim \text{IG}(2, b_\phi)$. We use the 1990 scallop data to provide a prior mean for each covariance parameter, thus determining, for example, $b_{\sigma^2} = 1/\text{E}(\sigma^2)$. The prior mean guesses are point estimates obtained by fitting (2.4) to the Matheron semivariogram of the 1990 data by using any convenient algorithm described in Cressie (1993). Finally, $\mu \sim \text{N}(a_\mu, b_\mu)$, where a_μ is the 1990 log scaled mean of 3.5 and $b_\mu = 1$. For the Bessel mixtures, we must add a prior for either $\mathbf{w}$ or $\boldsymbol{\phi}$. We use constant priors in both cases, assuming $\mathbf{w}$ to be Dirichlet($\alpha_1 = 1, \alpha_2 = 1, \ldots, \alpha_p = 1$) and $\boldsymbol{\phi}$ to be $\phi_{\max}$ times an ordered Dirichlet($\alpha_1 = 1, \alpha_2 = 1, \ldots, \alpha_p = 1$).

The resulting posterior distribution of $\boldsymbol{\theta}$ is only given up to proportionality as $f(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{y})\pi(\boldsymbol{\theta})$ and is not a standard form. Instead, to investigate features of the posterior, the duality between a density and samples from that density is utilized. Any desired attribute of the distribution can be obtained to arbitrary accuracy by sampling from the density. Because the dimensionality of $\boldsymbol{\theta}$ is low (at most nine in our examples), to sample the posterior we use a noniterative Monte Carlo method, employing an importance sampling density (ISD), say $g(\boldsymbol{\theta})$. Once $g(\boldsymbol{\theta})$ is chosen, we draw $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_V$ from $g(\boldsymbol{\theta})$ and form weights $\nu_i = [f(\boldsymbol{\theta}_i)]/[g(\boldsymbol{\theta}_i)]$. Let $\nu = \sum_{i=1}^V \nu_i$ and $q_i = \nu_i/\nu$, $i = 1, \ldots, V$. Monte Carlo integration for any posterior expectation, say $\text{E}(b(\boldsymbol{\theta})|\mathbf{y})$, takes the form $\sum q_i b(\boldsymbol{\theta}_i)$, while resampling the $\boldsymbol{\theta}_i$ using the probabilities q_i provides an approximate sample from the posterior (Smith and Gelfand 1992).

West's adaptive mixture method (1993) is used to iteratively construct a mixture distribution to use as a $g(\boldsymbol{\theta})$. Because West's procedure employs mixtures of multivariate normal distributions, all parameters should be transformed to the real line to improve the resulting $g(\boldsymbol{\theta})$. Starting from some $g^0(\boldsymbol{\theta}) = \text{N}(\boldsymbol{\theta}_0, \Sigma_0)$, draw $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_V$ from $g^0(\boldsymbol{\theta})$

and form the weights $\nu_i^0 = [f(\boldsymbol{\theta}_i)]/[g(\boldsymbol{\theta}_i)]$ to calculate

$$\bar{\boldsymbol{\theta}}_1 = \frac{\sum_{i=1}^V \nu_i^0 \boldsymbol{\theta}_i}{\sum_{i=1}^V \nu_i^0} \quad \text{and} \quad \Sigma_1 = \frac{\sum_{i=1}^V \nu_i^0 (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}}_1)(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}}_1)'}{\sum_{i=1}^V \nu_i^0},$$

and let $g^1(\boldsymbol{\theta}) = \sum_{i=1}^V \nu_i^0 \mathrm{N}(\boldsymbol{\theta}_i, \sum_1)$. Now repeat the process using $g^1(\boldsymbol{\theta})$ to form $g^2(\boldsymbol{\theta})$. When small variability in the weights ν_i is achieved, we stop. In practice, a few iterations usually provide a reasonable ISD for $f(\boldsymbol{\theta})$. Notice that in developing g and in evaluating the ν_i, we require many evaluations of $f(\boldsymbol{\theta})$ [i.e., many calculations of $|\Sigma(\boldsymbol{\alpha})|$ and $\Sigma(\boldsymbol{\alpha})^{-1}$].

5. MODEL CHOICE

General discussion of the model determination issue with regard to variogram specification appears to be virtually ignored in the literature. Penalized likelihood techniques such as the Akaike Information Criterion (Akaike 1973) and the Schwartz or Bayes Information Criterion (Schwartz 1978) can be used if $\mathbf{y}$ is parametrically specified. Otherwise, Webster and McBratney (1989) proposed, within a class of variogram models such as the Gaussian, minimizing a penalized residual sum of squares criterion to an arbitrarily lagged empirical variogram such as (2.2). However, for most variogram modeling applications, geoscientists seem content to use empirical wisdom to propose a suitable class of variogram models. Then, adopting a goodness-of-fit criterion, they obtain the best member of the class. There is no comparison across variogram classes. Is an exponential form better than a Gaussian? Is the flexible mixture of Bessel functions class substantially better than a Cauchy? In addressing this problem three issues emerge. What is an appropriate goodness-of-fit criterion? How do we penalize a variogram model for complexity? How do we adjust our model adequacy notion in the context of fitting variograms to data with enormous noise and weak signal? We clarify and answer these questions in this section.

Indexing models by m, with model m having parameters $\boldsymbol{\theta}^{(m)} = (\mu^{(m)}, \tau^{2(m)}, \sigma^{2(m)}, \phi^{(m)})$, the formal Bayesian approach is to compute the marginal density of the data evaluated at the observed $\mathbf{y}$. Unfortunately, this quantity can be difficult to compute and, moreover, it is only interpretable if the prior on $\boldsymbol{\theta}^{(m)}$ is proper. Additionally, even under a proper prior, this criterion arises from a hypothesis testing form of utility (Kadane and Dickey 1980) that does not, in a practical sense, reflect our utility for a variogram model.

Recall that our focus is explanation of spatial dependence, not spatial prediction. There is no elaboration of the mean structure; it is constant for all proposed models. Under the intrinsic hypothesis, spatial association is captured in the data by the set of $z_{ij} = (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2$. That is, under model m, $\mathrm{E}(z_{ij}|\boldsymbol{\theta}^{(m)}) = \eta_{ij}^{(m)} = 2\gamma(d_{ij}, \boldsymbol{\theta}^{(m)})$. Hence, model performance is reflected in the fit of the variogram to the set of z_{ij}. Because under (2.3), $z_{ij} \sim \eta_{ij}\chi_1^2 = \mathrm{Gamma}(1/2, 1/2\eta_{ij})$, variogram models elaborate mean structures for the z_{ij}'s.

To compare variograms we must work in the predictive space of the z_{ij}'s. (Comparison of $\boldsymbol{\theta}^{(m)}$'s makes no sense!) Hence, we need to specify a utility or, equivalently, a loss (negative utility) for taking action $a \in R^+$ when z is realized. A natural loss to

use is the deviance associated with the Gamma distribution, which, from McCullagh and Nelder (1989, p. 34), is

$$L(z,a) = \frac{z-a}{a} - \log\left(\frac{z}{a}\right). \tag{5.1}$$

Note that $L(z,a)$ takes its minimum at $a = z$ and increases strictly as a moves away from z in either direction.

Following Gelfand and Ghosh (in press), we extend (5.1) to a so-called balanced loss function (Zellner 1994):

$$L(z_{\text{rep}}, a; z_{\text{obs}}) = L(z_{\text{rep}}, a) + kL(z_{\text{obs}}, a). \tag{5.2}$$

In (5.2), z_{rep} is viewed as an independent replication of z_{obs} (i.e. if $z_{ij,\text{obs}}|\boldsymbol{\theta}^{(m)} \sim \eta_{ij}^{(m)}\chi_1^2$, then so is $z_{ij,\text{rep}}$). Then (5.2) views a as a compromise action that rewards closeness to z_{rep} but also penalizes for being too far from z_{obs}. The weight k indicates relative regret for departure from z_{obs} compared with departure from z_{rep}. Model choice is usually insensitive to specification of k, and we use $k = 1, 3, 9$ for illustration in Section 6.

We convert (5.2) to a model choice criterion by computing the sum over the $[N(N-1)/2]$ $z_{ij,\text{rep}}$'s,

$$D_{k,m} = \sum_{i,j} \min_{a_{ij}} \left\{ \mathrm{E}(L(z_{ij,\text{rep}}, a_{ij})) + kL(z_{ij,\text{obs}}, a_{ij}) \right\}. \tag{5.3}$$

In (5.3), the expectation is with respect to the predictive distribution of $z_{ij,\text{rep}}$ given $\mathbf{y}$ and the model m. That is, under model m, for each (i,j) pair we choose a_{ij} to minimize the expected posterior predictive loss under the distribution $f(z_{ij,\text{rep}}|\mathbf{y}, m)$. We then choose the model that makes $D_{k,m}$ the smallest, thus selecting the model with the largest expected utility. (One might criticize the additive form in (5.3) because the z_{ij} are not independent. However, it is simple and convenient, and the complexity of the joint dependence structure makes specification of an alternative form difficult.)

Fortunately, the minimization in (5.3) can be done explicitly under (5.1). After a bit of manipulation, we obtain

$$\begin{aligned} D_{k,m} &= (k+1)\sum_{i,j}\left\{\log\left(\frac{\lambda_{ij}^{(m)} + kz_{ij,\text{obs}}}{k+1}\right) - \frac{\log(\lambda_{ij}^{(m)}) + k\log(z_{ij,\text{obs}})}{k+1}\right\} \\ &\quad + \sum_{i,j}\left(\log(\lambda_{ij}^{(m)}) - \mathrm{E}\left(\log(z_{ij,\text{rep}}|\mathbf{y},m)\right)\right), \end{aligned} \tag{5.4}$$

where $\lambda_{ij}^{(m)} = \mathrm{E}(z_{ij,\text{rep}}|\mathbf{y}, m)$. The concavity of the log function ensures that both summations on the right hand side of (5.4) are positive. (As an aside, in theory $z_{ij,\text{obs}} > 0$ a.s., but in practice we may observe some $z_{ij} = 0$ as with, for example, the log counts in the scallop data example. A correction is needed and can be achieved by adding ϵ to $z_{ij,\text{obs}}$ where, say, ϵ is one-half of the smallest possible positive $z_{ij,\text{obs}}$).

The first term in (5.4), henceforth $G_{k,m}$, can be viewed as a goodness-of-fit piece. It would be 0 if each prediction for $z_{ij,\text{rep}}$, $\lambda_{ij}^{(m)}$, equaled $z_{ij,\text{obs}}$. The second term can be viewed as a penalty for model complexity, henceforth P_m. That is, expanding

$\log(z_{ij,\text{rep}})$ around $\log(\lambda_{ij}^{(m)})$ to second order and taking expectations, we find $\log(\lambda_{ij}^{(m)}) - \text{E}(\log(z_{ij,\text{rep}}|\mathbf{y}, m) \approx \text{var}(z_{ij,\text{rep}}|\mathbf{y}, m)/(\lambda_{ij}^{(m)})^2$. For underfitted models, predictive variances will tend to be large, hence, so will P_m. But also, for overfitted models, we expect inflated predictive variances, again making P_m large. Models that are too simple will do poorly in both $G_{k,m}$ and P_m. As the variogram model becomes increasingly complex, we anticipate a trade-off; $G_{k,m}$ will decrease but eventually P_m will increase. Thus, complexity is discouraged and parsimony is encouraged. Here, we find ourselves in the same spirit as familiar penalized likelihood approaches (e.g., Akaike 1973; Schwartz 1978), but for us the penalty term falls out as a result of the utility maximization.

The criterion in (5.4) is readily computed using Monte Carlo integration. If $\mathbf{z}_{\text{rep}}$ is the vector of $z_{ij,\text{rep}}$'s then

$$f(\mathbf{z}_{\text{rep}}|\mathbf{y}, m) = \int f(\mathbf{z}_{\text{rep}}|\boldsymbol{\theta}^{(m)})\ f(\boldsymbol{\theta}^{(m)}|\mathbf{y}, m)\ d\boldsymbol{\theta}^{(m)}. \tag{5.5}$$

The simulation-based model-fitting approach enables routine sampling of (5.5). Suppose $\boldsymbol{\theta}_\ell^{(m)^*}$, $\ell = 1, \ldots, L$ is a sample from the posterior $f(\boldsymbol{\theta}^{(m)}|\mathbf{y}, m)$ and that for each $\boldsymbol{\theta}_\ell^{(m)^*}$ we draw $\mathbf{y}_\ell^*$ according to (2.3). Transforming $\mathbf{y}_\ell^*$ to $\mathbf{z}_\ell^*$ immediately yields a random draw from $f(\mathbf{z}_{\text{rep}}|\mathbf{y}, m)$. But then the set of $z_{ij,\ell}^*$ is a sample from which we can compute a Monte Carlo integration for $\lambda_{ij}^{(m)}$ (i.e., $\sum z_{ij,\ell}^*/L$) and for $\text{E}(\log(z_{ij,\text{rep}}|\mathbf{y}, m))$ [i.e., $\sum \log(z_{ij,\ell}^*)/L$].

Finally, we note that any plausible variogram model will provide an inadequate fit to the $z_{ij,\text{obs}}$. In the simplest sense, we are fitting a low-dimensional parametric variogram model (perhaps 10 parameters at most) to $N(N-1)/2$ data points. (For the 1993 scallop data in Section 6, $N = 147$ so $N(N-1)/2 = 10,731$). This is evident from the semivariogram cloud of the $(d_{ij}, (1/2)z_{ij,\text{obs}})$ pairs in Figure 3(a).

The situation is analogous to simple linear regression with levels, say $x_1, x_2, \ldots, x_R$, and independent replications $y_{ij}, j = 1, \ldots, n_i$ at x_i. Letting $\hat{\mu}_i^{(m)}$ be the fitted value for $\text{E}(y_{ij})$ under model m, the error sum of squares is factored as

$$\sum_{i,j}(y_{ij} - \hat{\mu}_i^{(m)})^2 = \sum_i n_i(\bar{y}_i - \hat{\mu}_i^{(m)})^2 + \sum_{i,j}(y_{ij} - \bar{y}_i)^2. \tag{5.6}$$

The first term on the right hand side of (5.6) measures the lack-of-fit of model m by comparison with $\bar{y}_i$, the sample guess for the signal $\text{E}(y_{ij})$. The second term measures pure error, capturing the noise (variability) of the y_{ij} about $\text{E}(y_{ij})$. Typically, the second term dominates, but the first is more informative for model comparison.

By analogy, when looking at the set of $z_{ij,\text{obs}}$, we customarily create a lagged empirical semivariogram such as the Matheron estimator in (2.2). The γ_r^*'s that define this estimator play the role of the $\bar{y}_i$'s in (4.6). Goodness-of-fit of the semivariogram model is considered with respect to the empirical semivariogram. The semivariogram cloud of $N(N-1)/2$ points is reduced to R points where R is the number of sets B_r.

Three remarks are needed at this point. First, unlike the regression problem, here the choice of R is arbitrary. Which empirical variogram should we reduce to? Second, again distinct from the regression problem, the set of $z_{ij,\text{obs}}$ with $d_{ij} \in B_r$ are not independent and more importantly are not symmetrically distributed about their mean. In fact, up to

a scale parameter, they follow a χ^2_1 distribution rather than a symmetric normal. Figure 3(b) shows a boxplot of the $z_{ij,\text{obs}}/2$ within each B_r for the scallop data and reveals their substantial skewness. Why should γ^*_r, the sample mean of the $z_{ij,\text{obs}}$, be a suitable center against which fit is judged? Third, Bayesian (or likelihood) model fitting conditions on $\mathbf{y}$ not $\mathbf{z}$. There is no attempt to find $\boldsymbol{\theta}^{(m)}$ to provide a good fit to the $z_{ij,\text{obs}}$'s. We should not expect a Bayesian analysis to provide a "best-fitting" variogram to an arbitrarily selected empirical variogram.

Returning to $D_{k,m}$, how can we separate signal from noise analogous to (5.6) in the regression problem? Suppose, for a given set $B_1, B_2, \ldots, B_R$, we define the function

$$G_k(c_1, c_2, \ldots, c_R) = \sum_{r=1}^{R} \sum_{\{(i,j):d_{ij}\in B_r\}} \left\{ \log\left(\frac{c_r + k z_{ij,\text{obs}}}{k+1}\right) - \frac{\log(c_r) + k\log(z_{ij,\text{obs}})}{k+1} \right\}. \quad (5.7)$$

For a given r, consider minimizing the inner summation in (5.7), which we define as $\zeta_k(c_r)$. Clearly, $\zeta_k(c_r) \to \infty$ as $c_r \to 0$ and as $c_r \to \infty$. Also, because

$$\frac{c_r \zeta_k'(c_r)}{N_{B_r}} = \frac{1}{N_{B_r}} \sum_{\{(i,j):d_{ij}\in B_r\}} \frac{c_r}{c_r + k z_{ij,\text{obs}}} - \frac{1}{k+1} \in \left(-\frac{1}{k+1}, \frac{k}{k+1}\right)$$

and increases in c_r, the minimizing value, $\hat{c}_r$, is unique and can be calculated by any convenient root-finding algorithm. Then, the set $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_R$ provides the smallest possible value of $G_{k,m}$ when using a constant c_r over B_r and thus provides an empirical variogram arising from (5.4).

We may argue that, because the z_{ij}'s follow scaled χ^2_1 distributions, the $\hat{c}_r$'s provide a more natural variogram estimator than, say, the $2\gamma^*_r$'s in (2.2). We define this new empirical variogram to be the *deviance* variogram. It is data-based, motivated only by the presumed normality of $\mathbf{y}$, not by any model for γ. Because

$$\frac{\eta_r \zeta_k'(c_r)}{N_{B_r}} \approx \mathrm{E}\left(\frac{c_r}{c_r + kz} | z \sim \eta_r \chi^2_1\right) - \frac{1}{k+1},$$

if, for instance $k = 1$, then

$$\frac{\eta_r \zeta_1'(\eta_r)}{N_{B_r}} \approx \mathrm{E}\left((1+\omega)^{-1} | \omega \sim \chi^2_1\right) - \frac{1}{2} = .152$$

while

$$\frac{.385\eta_r \zeta_k'(.385\eta_r)}{N_{B_r}} \approx \mathrm{E}\left(.385(.385+\omega)^{-1} | \omega \sim \chi^2_1\right) - \frac{1}{2} = 0.$$

Hence, we expect $\hat{c}_r (\approx .385\eta_r) < 2\gamma^*_r (\approx \eta_r)$; that is, that the deviance empirical semivariogram will lie below the Matheron estimator. In Figure 4, we overlay the deviance empirical semivariogram for $k = 1$ on the Matheron estimate for the 1993 scallop data using lag $\delta = .05$ and observe this to be the case. Lastly, analogous to (5.6), we can subtract $G_k(\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_R)$ from $G_{k,m}$ to obtain $G_{k,m,\text{adj}}$, an adjusted goodness-of-fit value.

This value reflects the lack of fit of the variogram model m relative to the best fitting empirical variogram using $B_1, B_2, \ldots, B_R$.

6. EXAMPLE

The dataset of scallop catches in the Atlantic Ocean mentioned in Section 1 is examined for spatial correlation. After some preliminary exploratory discussion, we examine the results of parametric and nonparametric variogram model-fitting to these data. After choosing the best fitting model of each type, we summarize resultant inference for each. We note that the sites are given as (latitude,longitude) coordinates. In the New York Bight region, 1° latitude $\neq$ 1° longitude, so Euclidean distances using these coordinates do not agree with physical distances. Correction could be made using planar projections.

6.1 DATA AND EXPLORATORY TECHNIQUES

For the 1993 scallop data, Figure 3(a) shows the semivariogram cloud and Figure 3(b) shows boxplots formed from the cloud using the arbitrary lag $\delta = .05$. The 10,731 pairs of points that produce the semivariogram cloud do not reveal any distinct pattern. However, the boxplots and the Matheron and deviance empirical semivariograms (see Sec. 5), each based on lag $\delta = .05$ (Fig. 4), clearly exhibit spatial dependence in the sense that when separation distances are small, the spatial variability tends to be less.

For the choice of $\phi_{\max}$ in the nonparametric setup, we selected seven sign changes or three Bessel periods. With $d_{ij}^{\max} = 2.83$ degrees, $\phi_{\max}$ becomes 7.5.

A sensitivity analysis with two Bessel mixtures ($p = 2$) having a fixed weight w_1 and random nodes was undertaken. Two, four, and five Bessel periods revealed little difference in results as compared with three. However, when one Bessel period was examined ($\phi_{\max} = 3$), the model fit poorly and in fact ϕ_p was just smaller than 3. This is an indication that more flexibility (a larger value of $\phi_{\max}$) is required.

6.2 FITTED SEMIVARIOGRAM MODELS

All of the parametric models of Table 1 and nonparametric Bessel mixtures with different combinations of fixed and random parameters were fitted to the 1993 scallop data. Figure 5 shows the posterior mean of each respective semivariogram, and Table 2 provides the model choice criteria for each model along with the independence model $[\Sigma(\boldsymbol{\alpha}) = (\tau^2 + \sigma^2)I]$. The simulation error associated with the entries in Table 2 can most easily be assessed by repeated fitting of the models. In doing so, we conclude that a 3σ range is at most 30. All variogram models fit better than the independence model, which again supports the presence of spatial correlation. Of the parametric models, the Cauchy and Gaussian fit best using the $D_{k,m}$ and G_k criteria. Taking $k = 1$, the value of $G_1(\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_R)$ associated with lag $\delta = .05$ is 8961.6. For other lags, $\delta = .1$ and $\delta = .025$, the values are 8971.1 and 8942.8, respectively, confirming the level of noise about the deviance variogram. The $G_{1,m,\text{adj}} = G_{1,m} - G_1(\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_R)$ value in Table 2 is the sharpened goodness-of-fit criterion. The variogram model with the smallest $G_{1,m,adj}$ value (the nonparametric mixture of five Bessel functions with random weights and fixed ϕ's) is the best fitting variogram model; however, the model with the smallest penalty component, P_m, is the Cauchy. The wide range of estimated (posterior mean) variograms in Figure 6 indicates considerable sensitivity to the choice of the parametric form.

Table 2. Model Choice for Fitted Variogram Models

Name	$G_{1,m}$	$G_{1,m,adj}$	P_m	$D_{1,m}$	$D_{3,m}$	$D_{9,m}$
Parametric:						
Exponential	10,961.6	2,000.0	13,897.7	24,859.3	40,905.7	75,202.7
Gaussian	10,860.5	1,898.9	13,843.2	24,703.7	40,519.6	74,317.2
Cauchy	10,695.6	1,734.0	13,808.5	24,504.1	39,950.1	72,940.2
Spherical	11,450.0	2,488.4	13,965.7	25,415.7	42,381.7	78,665.1
Bessel	11,042.2	2,080.6	14,032.5	25,074.7	41,291.0	75,972.3
Independent	11,577.2	2,615.6	16,159.3	27,736.5	44,824.7	81,426.4
Semiparametric:						
Fixed ϕ_ℓ's/Random w_ℓ's						
Two	11,072.8	2,111.2	13,967.5	25,040.3	41,320.5	76,133.8
Three	10,593.7	1,632.1	13,818.6	24,412.3	39,725.0	72,436.5
Four	10,934.4	1,972.8	13,870.2	24,804.6	40,809.8	75,026.9
Five	10,564.5	1,602.9	13,817.5	24,382.0	39,627.6	72,192.5
Random ϕ_ℓ's/Fixed w_ℓ's						
Two	10,677.0	1,715.4	13,906.3	24,583.3	40,091.9	73,236.0
Three	10,685.0	1,723.4	13,968.6	24,653.6	40,181.3	73,364.9
Four	10,645.8	1,684.2	13,915.7	24,561.5	40,000.3	72,989.3
Five	10,614.3	1,652.7	13,902.7	24,517.0	39,887.3	72,725.1

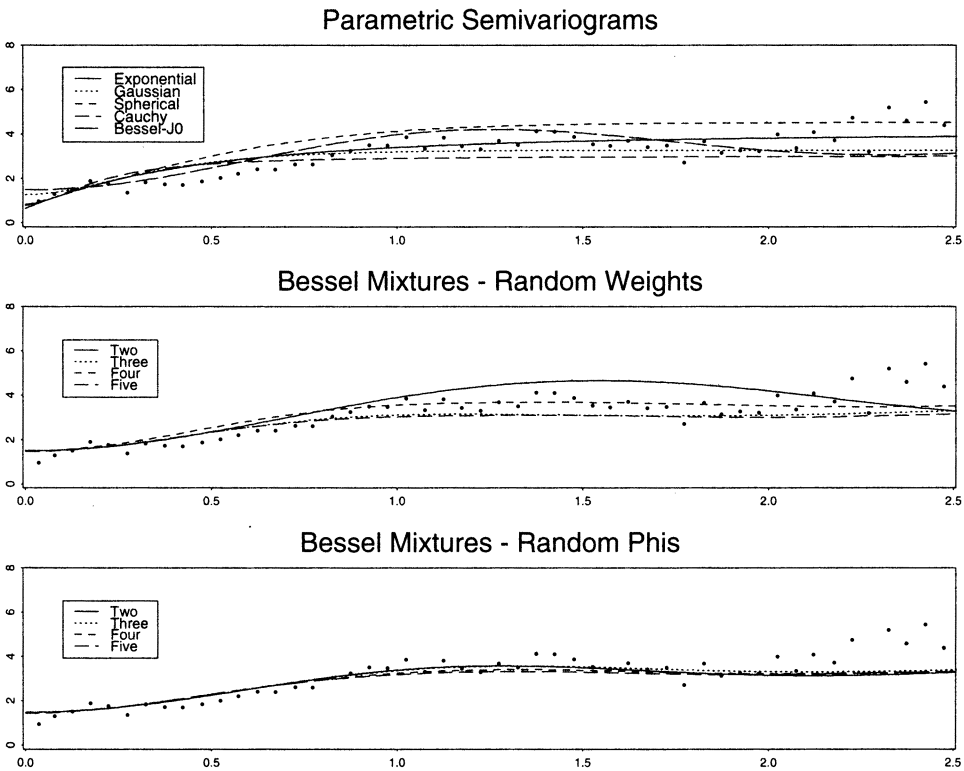

Figure 6. Posterior Means for Various Semivariogram Models.

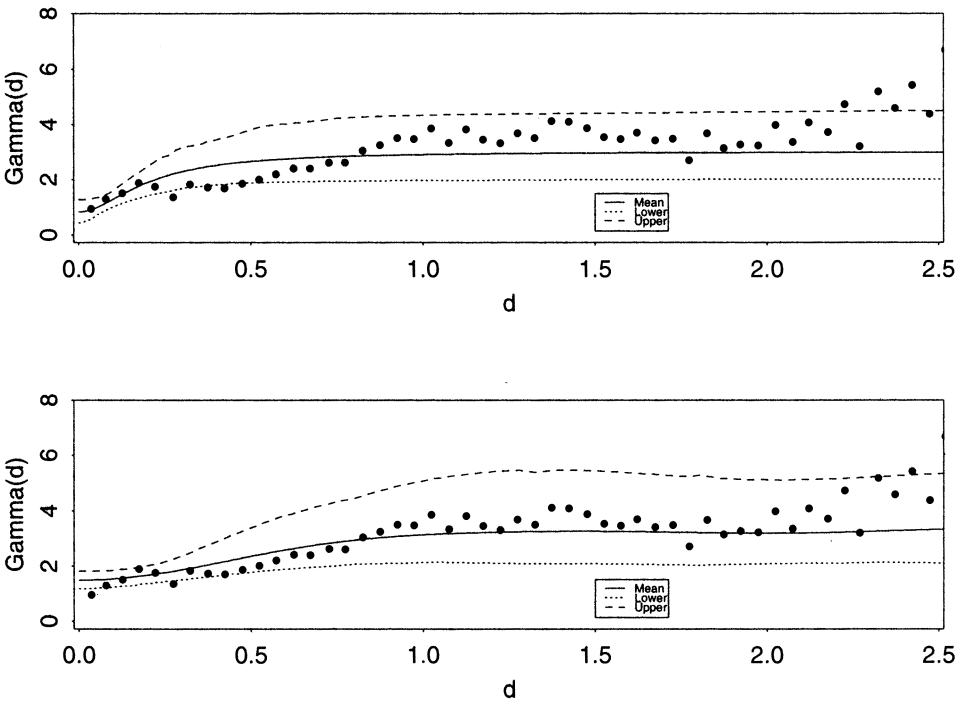

Figure 7. Interval Estimates for Cauchy and the Five-Component Bessel Mixture With Fixed ϕ's. Matheron empirical semivariogram (•) at lag δ = .05 shown for reference.

Of the Bessel mixtures, the five-component model with random weights and fixed ϕ's fit best where, given $\phi_{\max} = 7.5$, the nodes were fixed to be $\phi_1 = 1.25, \phi_2 = 2.5, \phi_3 = 3.75, \phi_4 = 5$, and $\phi_5 = 6.25$. One would expect that the fit measured by the $G_{k,m}$ criterion should improve with increasing p. However, the models are not nested by p except, for instance, the $p = 2$ model, which is a special case of the $p = 5$ model. Thus, it can occur that the four-component fixed ϕ model performs worse than the three-component model. The random ϕ Bessel mixture models were all very close and, as a class, these models fit as well as or better than the best parametric model. Hence, modeling mixtures of Bessel functions appears more sensitive to the choice of fixed ϕ's than to fixed weights.

6.3 THE BEST-FITTING PARAMETRIC AND NONPARAMETRIC MODELS

In this section, the best-fitting parametric model (Cauchy) and nonparametric model (mixture of five Bessel functions with random weights and fixed ϕ's) are further examined. Figure 7 displays the posterior means for each with 95% interval estimates. The Bessel mixture model appears to align with the Matheron estimator better than the Cauchy except near the origin. Point and interval estimates for all parameters are presented in Table 3. As seen in Figure 7 and also from the interval estimates for their respective sills, the Cauchy model has a tighter upper confidence bound for its variogram. Also, nega-

Table 3. Summary Statistics From the Posterior by Model

Parameter	*Posterior Mean*	*95% Interval Estimate*
Cauchy:		
Nugget(τ^2)	.84	(.44, 1.28)
σ^2	2.17	(1.18, 3.70)
ϕ	29.41	(6.71, 91.09)
μ	3.02	(2.25, 3.99)
Sill	3.01	(2.03, 4.52)
r_C	.93	(.46, 1.68)
r_V	.78	(.41, 1.39)
Bessel:		
Nugget(τ^2)	1.47	(1.18, 1.81)
σ^2	1.59	(.69, 3.68)
μ	2.86	(1.92, 3.78)
Sill	3.05	(2.08, 4.78)

tive covariances among the α parameters of the Cauchy model arise in agreement with intuition. That is, if ϕ were held constant in the Cauchy model and σ^2 were increased, then τ^2 must decrease to accommodate the data. Likewise, if τ^2 were held constant and σ^2 increased, then the range must also increase, forcing ϕ to decrease.

For the Cauchy model, posterior distributions for the τ^2, σ^2, μ, and the sill are each represented by the solid line in Figures 8(a–d), respectively. Figure 8(e) shows ϕ for the Cauchy correlation structure, and the ranges r_C (solid line) and r_V (dashed line) for the Cauchy semivariogram are presented in Figure 8(f). The posterior means for the ranges r_C and r_V are .93 and .78 degrees. The posteriors for τ^2, σ^2, μ, and the sill for the five-component random-weight Bessel model are given by the dotted line in Figures 8(a–d). The posterior medians for the five weights are .203, .271, .337, .016, and .173, respectively. The two models have very different nugget(τ^2), but their sills are nearly identical with the Bessel mixture being slightly more right skewed. Again, because the Bessel function is not monotone, the range is not defined.

7. CONCLUSIONS

We have proposed a fully Bayesian approach to fitting variogram models, arguing its advantages over customary methods. We have examined a broad range of models including an arbitrarily flexible class obtained through mixtures. We have shown how model determination can be carried out in this framework to investigate adequacy and comparison. We also have.demonstrated the scope of possible inference.

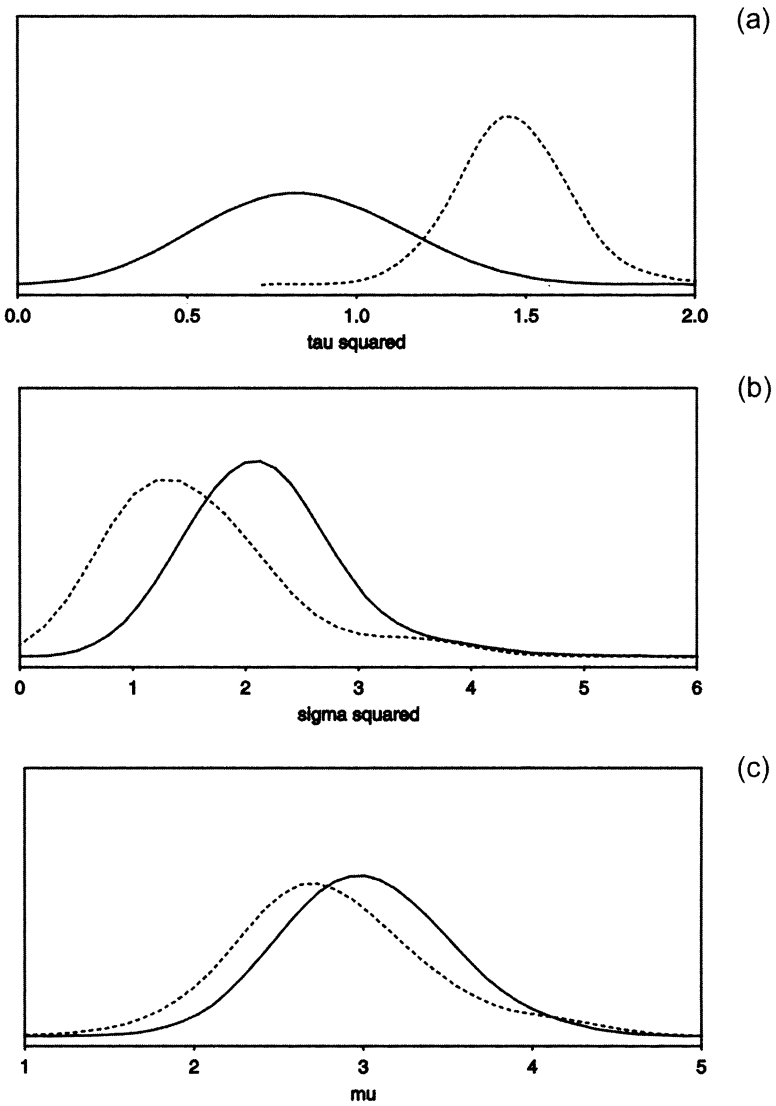

Figure 8. Posterior Distributions for τ^2 (a), σ^2 (b), μ (c) Under Cauchy Model (—) and Under the Mixture of Five-Bessels Model (- - -).

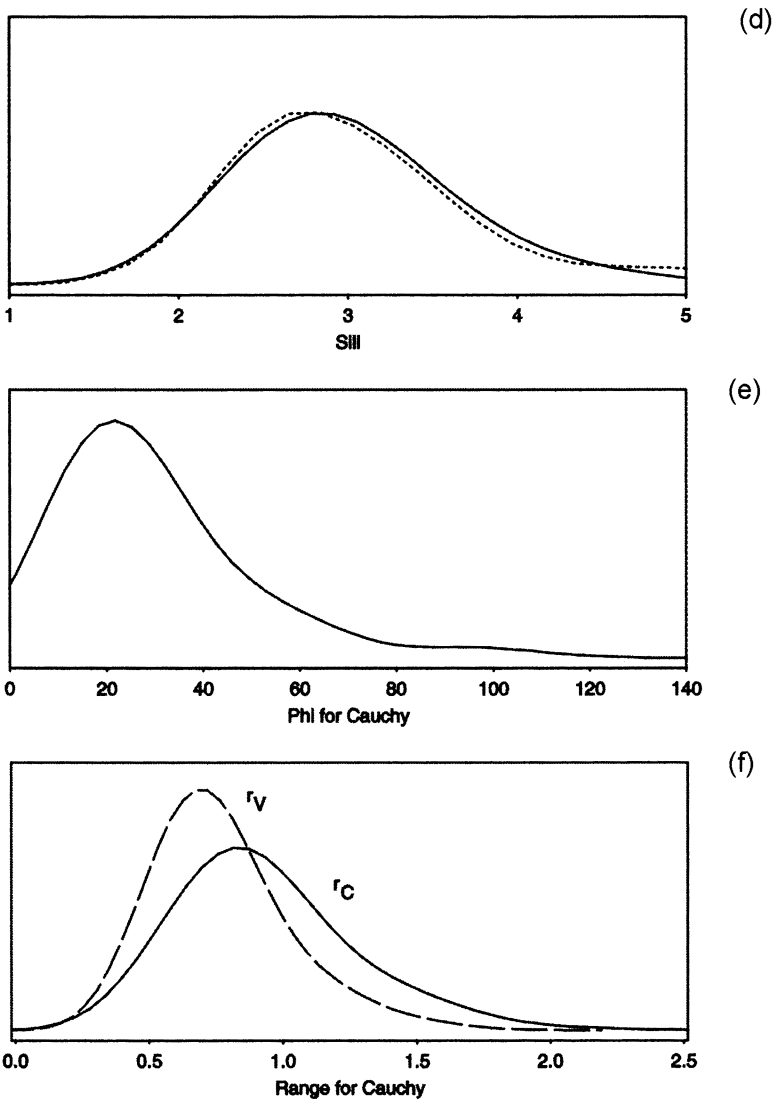

Figure 8. (continued) Posterior Distributions for sill (d) under Cauchy model (—) and Under the Mixture of Five-Bessels Model (- - -). Cauchy model ϕ (e) and ranges r_C (—) and r_V (– –) (f).

ACKNOWLEDGMENTS

The work of the second author was supported in part by NSF DMS 96-25383. The authors acknowledge the associate editor and three referees for insightful comments on an earlier version of the manuscript, which substantially improved the revision.

[Received January 1997. Revised July 1997.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B. Petrov and F. Csáki, Budapest: Akadémia Kiadó, pp. 267–281.

Armstrong, M., and Diamond, P. (1984), "Testing Variograms for Positive Definiteness," *Math Geology*, 24, 135–147.

Armstrong, M., and Jabin, R. (1981), "Variogram Models Must Be Positive Definite," *Math Geology*, 13, pp. 455–459.

Barry, R. P., and Ver Hoef, J. M. (1996), Blackbox Kriging: Spatial Prediction Without Specifying Variogram Models," *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 297–322.

Cherry, S., Banfield, J., and Quimby, W. (1996), "An Evaluation of a Nonparametric Method of Estimating Semivariograms of Isotropic Spatial Process," *Journal of Applied Statistics*, 23, 435–449.

Chung, K. (1974), *A Course in Probability Theory*, New York: Academic Press.

Christakos, G. (1984), "On the Problem of Permissible Covariance and Variogram Models," *Water Resources Research*, 20, 251–265.

Cressie, N. (1985), "Fitting Variogram Models by Weighted Least Squares," *Math Geology*, 17, 563–586.

——— (1993), *Statistics for Spatial Data*, New York: Wiley.

DeOliveira, V., Kedem, B., and Short, D. (1997), "Bayesian Prediction of Transformed Gaussian Random Fields," technical report, Department of Mathematics, University of Maryland, College Park, MD.

Diggle, P., Liang, K-L., and Zeger, S. (1994), *Analysis of Longitudinal Data*, New York: Oxford Science Publications.

Ecker, M., and Heltshe, J. (1994), "Geostatistical Estimates of Scallop Abundance," in *Case Studies in Biometry*, eds. N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhouse, New York: Wiley, pp. 107–124.

Feller, W. (1966), *An Introduction to Probability Theory and its Applications, Volume II*, New York: Wiley.

Gaudard, M., Karson, M., Linder, E., and Sinha, D. (1995), "Modeling Precipitation Using Bayesian Spatial Analysis," in *Proceedings of the Section on Bayesian Inference, American Statistical Association*, pp. 173–177.

Gelfand, A. E., and Ghosh, S. K. (in press), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85.

Hall, P., Fisher, N., and Hoffmann, B. (1994), "On the Nonparametric Estimation of Covariance Functions," *The Annals of Statistics*, 22, 2115–2134.

Handcock, M., and Stein, M. (1993), "A Bayesian Analysis of Kriging," *Technometrics*. 35, 403–410.

Handcock, M., and Wallis, J. (1994), "An Approach to Statistical Spatial-Temporal Modeling of Meteorological Fields (with discussion)," *Journal of the American Statistical Association*, 89, 368–390.

Journel, A., and Huijbregts, C. (1978), *Mining Geostatistics*, New York: Academic Press.

Journel, A., and Rossi, M. (1989), "When Do We Need a Trend Model in Kriging?" *Math Geology*. 21, 715–739.

Kadane, J., and Dickey, J. (1980), "Bayesian Decision Theory and the Simplification of Models," in *Evaluations of Econometric Models*, eds. J. Kmenta and J. Ramsey, New York: Academic Press, pp. 245–268.

Kaluzny, S., Vega, S., Cardosa, T., and Shelly, A. (1996), *S+SpatialStats User's Manual,* Seattle: MathSoft Inc.

Lele, S. (1995), "Inner Product Matrices, Kriging, and Nonparametric Estimation of Variogram," *Math Geology,* 27, 673–692.

Mantoglou, A., and Wilson, J. (1982), "The Turning Bands Methods for Simulation of Random Fields Using Line Generation by Spectral Method," *Water Resources Research,* 18, 1379–1394.

Matheron, G. (1963), "Principles of Geostatistics," *Economic Geology,* 58, 1246–1266.

McBratney, A., and Webster, R. (1986), "Choosing Functions for Semi-Variograms of Soil Properties and Fitting Them to Sampling Estimates," *Journal of Soil Science,* 37, 617–639.

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models,* New York: Chapman and Hall.

Sampson, P., and Guttorp, P. (1992), "Nonparametric Estimation of Nonstationary Spatial Covariance Structure. *Journal of the American Statistical Association,* 87, 108–119.

Schoenberg, I. (1938), "Metric Spaces and Completely Monotone Functions," *Annals of Mathematics,* 39, 811–841.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics,* 6, 461–464.

Shapiro, A., and Botha, J. (1991), "Variogram Fitting With a General Class of Conditionally Nonnegative Definite Functions," *Computational Statistics and Data Analysis,* 11, 87–96.

Smith, A. F. M., and Gelfand, A. E. (1992), "Bayesian Statistics Without Tears: A Sampling–Resampling Perspective," *The American Statistician,* 46, 84–88.

Webster, R., and McBratney, A. (1989), "On the Akaike Information Criterion for Choosing Models for Variograms of Soil Properties," *Journal of Soil Science,* 40, 493–496.

West, M. (1993), "Approximating Posterior Distributions by Mixtures," *Journal of the Royal Statistical Society,* Ser. B, 55, 563–586.

Yaglom, A. (1987), *Correlation Theory of Stationary and Related Random Functions I,* New York: Springer-Verlag.

Zellner, A. (1994), "Bayesian and Non-Bayesian Estimation Using Balanced Loss Functions," in *Statistical Decision Theory and Related Topics, V,* eds. S. Gupta and J. Berger, New York: Springer-Verlag, pp. 377–390.