good approximation to the intersite geodesic distances. This projection is useful for entering coordinates in spatial statistics software packages that require two-dimensional coordinate input and uses Euclidean metrics to compute distances (e.g., the variogram functions in `S+SpatialStats`, the `spatial.exp` function in `WinBUGS`, etc.).

(a) Compute the above projection for Chicago and Minneapolis ($N = 2$) and find the Euclidean distance between the projected coordinates. Compare with the geodesic distance. Repeat this exercise for New York and New Orleans.

(b) When will the above projection fail to work?

CHAPTER 2

Basics of point-referenced data models

In this chapter we present the essential elements of spatial models and classical analysis for point-referenced data. As mentioned in Chapter 1, the fundamental concept underlying the theory is a stochastic process $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$, where D is a fixed subset of r-dimensional Euclidean space. Note that such stochastic processes have a rich presence in the time series literature, where $r = 1$. In the spatial context, usually we encounter r to be 2 (say, northings and eastings) or 3 (e.g., northings, eastings, and altitude above sea level). For situations where $r > 1$, the process is often referred to as a *spatial process*. For example, $Y(\mathbf{s})$ may represent the level of a pollutant at site $\mathbf{s}$. While it is conceptually sensible to assume the existence of a pollutant level at all possible sites in the domain, in practice the data will be a partial realization of that spatial process. That is, it will consist of measurements at a finite set of locations, say $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, where there are monitoring stations. The problem facing the statistician is inference about the spatial process $Y(\mathbf{s})$ and prediction at new locations, based upon this partial realization.

This chapter is organized as follows. We begin with a survey of the building blocks of point-level data modeling, including stationarity, isotropy, and variograms (and their fitting via traditional moment-matching methods). We then add the spatial (typically Gaussian) process modeling that enables likelihood (and Bayesian) inference in these settings. We also illustrate helpful exploratory data analysis tools, as well as more formal classical methods, especially kriging (point-level spatial prediction). We close with short tutorials in `S+SpatialStats` and `geoR`, two easy to use and widely available point-level spatial statistical analysis packages.

The material we cover in this chapter is traditionally known as *geostatistics*, and could easily fill many more pages than we devote to it here. While we prefer the more descriptive term "point-level spatial modeling," we will at times still use "geostatistics" for brevity and perhaps consistency when referencing the literature.

2.1 Elements of point-referenced modeling

2.1.1 Stationarity

For our discussion we assume that our spatial process has a mean, say $\mu(s) = E(Y(s))$, associated with it and that the variance of $Y(\mathbf{s})$ exists for all $\mathbf{s} \in D$. The process $Y(\mathbf{s})$ is said to be *Gaussian* if, for any $n \geq 1$ and any set of sites $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$ has a multivariate normal distribution. The process is said to be *strictly stationary* if, for any given $n \geq 1$, any set of n sites $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and any $\mathbf{h} \in \Re^r$, the distribution of $(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))$ is the same as that of $(Y(\mathbf{s}_1 + \mathbf{h}), \ldots, Y(\mathbf{s}_n + \mathbf{h}))$. Here D is envisioned as $\Re^r$ as well.

A less restrictive condition is given by *weak stationarity* (also called second-order stationarity). Cressie (1993, p. 53) defines a spatial process to be weakly stationary if $\mu(\mathbf{s}) \equiv \mu$ (i.e., the process has a constant mean) and $Cov(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})) = C(\mathbf{h})$ for all $\mathbf{h} \in \Re^r$ such that $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$ both lie within D. (We note that, strictly speaking, for stationarity as a second-order property we will need only the second property; $E(Y(\mathbf{s}))$ need not equal $E(Y(\mathbf{s}+\mathbf{h}))$. But since we will apply the definition only to a mean 0 spatial residual term, this distinction is unimportant for us.) Weak stationarity implies that the covariance relationship between the values of the process at any two locations can be summarized by a covariance function $C(\mathbf{h})$, and this function depends only on the separation vector $\mathbf{h}$. Note that with all variances assumed to exist, strong stationarity implies weak stationarity. The converse is not true in general, but it *does* hold for Gaussian processes; see Exercise 2.

2.1.2 Variograms

There is a third type of stationarity called *intrinsic* stationarity. Here we assume $E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})] = 0$ and define

$$E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})]^2 = Var\,(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})) = 2\gamma(\mathbf{h})\ . \qquad (2.1)$$

Equation (2.1) makes sense only if the left-hand side depends *only* on $\mathbf{h}$ (so that the right-hand side can be written at all), and not the particular choice of $\mathbf{s}$. If this is the case, we say the process is *intrinsically stationary*. The function $2\gamma(\mathbf{h})$ is then called the *variogram*, and $\gamma(\mathbf{h})$ is called the *semivariogram*. (The covariance function $C(\mathbf{h})$ is sometimes referred to as the *covariogram*, especially when plotted graphically.) Note that intrinsic stationarity defines only the first and second moments of the differences $Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})$. It says nothing about the joint distribution of a collection of variables $Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)$, and thus provides no likelihood.

It is easy to see the relationship between the variogram and the covariance function:

$$\begin{aligned} 2\gamma(\mathbf{h}) &= Var\,(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})) \\ &= Var(Y(\mathbf{s}+\mathbf{h})) + Var(Y(\mathbf{s})) - 2Cov(Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s})) \\ &= C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h}) \\ &= 2\left[C(\mathbf{0}) - C(\mathbf{h})\right]\ . \end{aligned}$$

Thus,

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})\ . \qquad (2.2)$$

From (2.2) we see that given C, we are able to recover γ easily. But what about the converse; in general, can we recover C from γ? Here it turns out we need to assume a bit more: if the spatial process is *ergodic*, then $C(\mathbf{h}) \to 0$ as $||\mathbf{h}|| \to \infty$, where $||\mathbf{h}||$ denotes the length of the $\mathbf{h}$ vector. This is an intuitively sensible condition, since it means that the covariance between the values at two points vanishes as the points become further separated in space. But taking the limit of both sides of (2.2) as $||\mathbf{h}|| \to \infty$, we then have that $lim_{||\mathbf{h}||\to\infty}\gamma(\mathbf{h}) = C(\mathbf{0})$. Thus, using the dummy variable $\mathbf{u}$ to avoid confusion, we have

$$C(\mathbf{h}) = C(\mathbf{0}) - \gamma(\mathbf{h}) = lim_{||\mathbf{u}||\to\infty}\gamma(\mathbf{u}) \; - \; \gamma(\mathbf{h})\ . \qquad (2.3)$$

In general, the limit on the right-hand side need not exist, but if it does, then the process is weakly (second-order) stationary with $C(\mathbf{h})$ as given in (2.3). We then have a way to determine the covariance function C from the semivariogram γ. Thus weak stationarity implies intrinsic stationarity, but the converse is not true; indeed, the next section offers examples of processes that are intrinsically stationary but not weakly stationary.

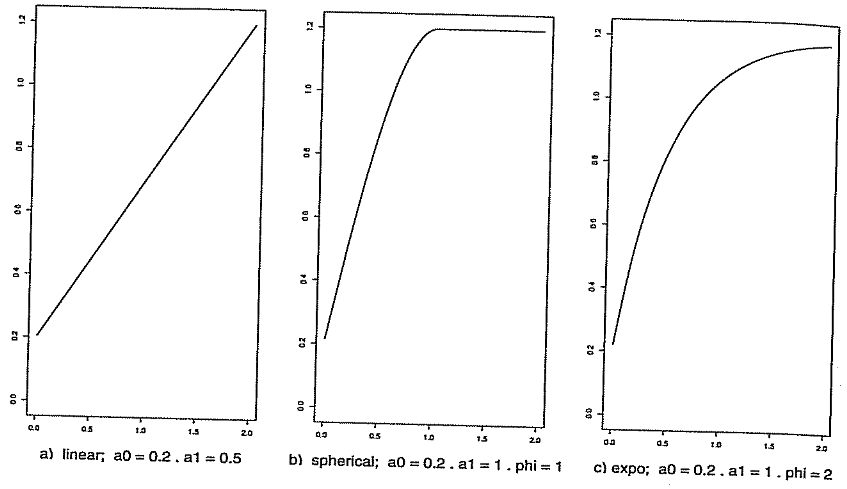
A valid variogram necessarily satisfies a negative definiteness condition. In fact, for any set of locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and any set of constants $a_1, \ldots, a_n$ such that $\sum_i a_i = 0$, if $\gamma(\mathbf{h})$ is valid, then

$$\sum_i \sum_j a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0\ . \qquad (2.4)$$

To see this, note that

$$\begin{aligned} \sum_i \sum_j a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) &= \frac{1}{2} E \sum_i \sum_j a_i a_j (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2 \\ &= -E \sum_i \sum_j a_i a_j Y(\mathbf{s}_i) Y(\mathbf{s}_j) \\ &= -E \left[\sum_i a_i Y(\mathbf{s}_i) \right]^2 \; \leq \; 0\ . \end{aligned}$$

Note that, despite the suggestion of expression (2.2), there is no relationship between this result and the positive definiteness condition for



a) linear; a0 = 0.2 . a1 = 0.5

b) spherical; a0 = 0.2 . a1 = 1 . phi = 1

c) expo; a0 = 0.2 . a1 = 1 . phi = 2

Figure 2.1 *Theoretical semivariograms for three models: (a) linear, (b) spherical, and (c) exponential.*

covariance functions (see Subsection 2.2.2). Cressie (1993) discusses further necessary conditions for a valid variogram. Lastly, the condition (2.4) emerges naturally in ordinary kriging (see Section 2.4).

2.1.3 Isotropy

Another important related concept is that of isotropy (as mentioned in Subsection 1.1.1). If the semivariogram function $\gamma(\mathbf{h})$ depends upon the separation vector only through its length $||\mathbf{h}||$, then we say that the process is *isotropic*; if not, we say it is *anisotropic*. Thus for an isotropic process, $\gamma(\mathbf{h})$ is a real-valued function of a univariate argument, and can be written as $\gamma(||\mathbf{h}||)$. If the process is intrinsically stationary and isotropic, it is also called *homogeneous*.

Isotropic processes are popular because of their simplicity, interpretability, and, in particular, because a number of relatively simple parametric forms are available as candidates for the semivariogram. Denoting $||\mathbf{h}||$ by t for notational simplicity, we now consider a few of the more important such forms.

1. *Linear:*

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 t & \text{if } t > 0,\ \tau^2 > 0,\ \sigma^2 > 0 \\ 0 & \text{otherwise} \end{cases} .$$

Note that $\gamma(t) \to \infty$ as $t \to \infty$, and so this semivariogram does not correspond to a weakly stationary process (although it is intrinsically stationary). This semivariogram is plotted in Figure 2.1(a) using the parameter values $\tau^2 = 0.2$ and $\sigma^2 = 0.5$.

2. *Spherical:*

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t \geq 1/\phi, \\ \tau^2 + \sigma^2 \left\{ \frac{3\phi t}{2} - \frac{1}{2}(\phi t)^3 \right\} & \text{if } 0 < t \leq 1/\phi, \\ 0 & \text{otherwise} \end{cases} .$$

The spherical semivariogram is valid in $r = 1, 2$, or 3 dimensions, but for $r \geq 4$ it fails to correspond to a spatial variance matrix that is positive definite (as required to specify a valid joint probability distribution). The spherical form does give rise to a stationary process and so the corresponding covariance function is easily computed (see the exercises that follow).

This variogram owes its popularity largely to the fact that it offers clear illustrations of the *nugget*, *sill*, and *range*, three characteristics traditionally associated with variograms. Specifically, consider Figure 2.1(b), which plots the spherical semivariogram using the parameter values $\tau^2 = 0.2$, $\sigma^2 = 1$, and $\phi = 1$. While $\gamma(0) = 0$ by definition, $\gamma(0^+) \equiv lim_{t \to 0^+} \gamma(t) = \tau^2$; this quantity is the *nugget*. Next, $\lim_{t \to \infty} \gamma(t) = \tau^2 + \sigma^2$; this asymptotic value of the semivariogram is called the *sill*. (The sill minus the nugget, which is simply σ^2 in this case, is called the *partial sill*.) Finally, the value $t = 1/\phi$ at which $\gamma(t)$ first reaches its ultimate level (the sill) is called the *range*. It is for this reason that many of the variogram models of this subsection are often parametrized through $R \equiv 1/\phi$. Confusingly, both R and ϕ are sometimes referred to as the *range* parameter, although ϕ is often more accurately referred to as the *decay* parameter.

Note that for the linear semivariogram, the nugget is τ^2 but the sill and range are both infinite. For other variograms (such as the next one we consider), the sill is finite, but only reached asymptotically.

3. *Exponential:*

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 (1 - \exp(-\phi t)) & \text{if } t > 0, \\ 0 & \text{otherwise} \end{cases} .$$

The exponential has an advantage over the spherical in that it is simpler in functional form while still being a valid variogram in all dimensions (and without the spherical's finite range requirement). However, note from Figure 2.1(c), which plots this semivariogram assuming $\tau^2 = 0.2$, $\sigma^2 = 1$, and $\phi = 2$, that the sill is only reached asymptotically, meaning that strictly speaking, the range $R = 1/\phi$ is infinite. In cases like this, the notion of an *effective range* if often used, i.e., the distance at which there is essentially no lingering spatial correlation. To make this notion precise, we must convert from γ scale to C scale (possible here since $\lim_{t \to \infty} \gamma(t)$ exists; the exponential is not only intrinsically but also weakly stationary).

From (2.3) we have

$$\begin{aligned} C(t) &= lim_{u\to\infty}\gamma(u) - \gamma(t) \\ &= \tau^2 + \sigma^2 - \left[\tau^2 + \sigma^2(1 - \exp(-\phi t))\right] \\ &= \sigma^2 \exp(-\phi t)\ . \end{aligned}$$

Hence

$$C(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t = 0 \\ \sigma^2 \exp(-\phi t) & \text{if } t > 0 \end{cases} . \tag{2.5}$$

If the nugget $\tau^2 = 0$, then this expression reveals that the correlation between two points t units apart is $\exp(-\phi t)$; note that $\exp(-\phi t) = 1^-$ for $t = 0^+$ and $\exp(-\phi t) = 0$ for $t = \infty$, both in concert with this interpretation.

A common definition of the *effective range*, t_0, is the distance at which this correlation has dropped to only 0.05. Setting $\exp(-\phi t_0)$ equal to this value we obtain $t_0 \approx 3/\phi$, since $\log(0.05) \approx -3$. The range will be discussed in more detail in Subsection 2.2.2.

Finally, the form of (2.5) gives a clear example of why the nugget (τ^2 in this case) is often viewed as a "nonspatial effect variance," and the partial sill (σ^2) is viewed as a "spatial effect variance." Along with ϕ, a statistician would likely view fitting this model to a spatial data set as an exercise in estimating these three parameters. We shall return to variogram model fitting in Subsection 2.1.4.

4. *Gaussian:*

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2\left(1 - \exp\left(-\phi^2 t^2\right)\right) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} . \tag{2.6}$$

The Gaussian variogram is an analytic function and yields very smooth realizations of the spatial process. We shall say much more about process smoothness in Subsection 2.2.3.

5. *Powered exponential:*

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2\left(1 - \exp\left(-|\phi t|^p\right)\right) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} . \tag{2.7}$$

Here $0 < p \leq 2$ yields a family of valid variograms. Note that both the Gaussian and the exponential forms are special cases of this one.

6. *Rational quadratic:*

$$\gamma(t) = \begin{cases} \tau^2 + \frac{\sigma^2 t^2}{(1+\phi t^2)} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} .$$

7. *Wave:*

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2\left(1 - \frac{\sin(\phi t)}{\phi t}\right) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} .$$

Model	Covariance function, $C(t)$
Linear	$C(t)$ does not exist
Spherical	$C(t) = \begin{cases} 0 & \text{if } t \geq 1/\phi \\ \sigma^2\left[1 - \frac{3}{2}\phi t + \frac{1}{2}(\phi t)^3\right] & \text{if } 0 < t \leq 1/\phi \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Exponential	$C(t) = \begin{cases} \sigma^2 \exp(-\phi t) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Powered exponential	$C(t) = \begin{cases} \sigma^2 \exp(-\lvert\phi t\rvert^p) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Gaussian	$C(t) = \begin{cases} \sigma^2 \exp(-\phi^2 t^2) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Rational quadratic	$C(t) = \begin{cases} \sigma^2\left(1 - \frac{t^2}{(1+\phi t^2)}\right) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Wave	$C(t) = \begin{cases} \sigma^2 \frac{\sin(\phi t)}{\phi t} & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Power law	$C(t)$ does not exist
Matérn	$C(t) = \begin{cases} \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(2\sqrt{\nu}t\phi\right)^\nu K_\nu(2\sqrt{\nu}t\phi) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Matérn at $\nu = 3/2$	$C(t) = \begin{cases} \sigma^2\left(1 + \phi t\right)\exp\left(-\phi t\right) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$

Table 2.1 *Summary of covariance functions (covariograms) for common parametric isotropic models.*

Note this is an example of a variogram that is not monotonically increasing. The associated covariance function is $C(t) = \sigma^2 \sin(\phi t)/(\phi t)$. Bessel functions of the first kind include the wave covariance function and are discussed in detail in Subsections 2.2.2 and 5.1.3.

8. *Power law*

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 t^\lambda & \text{of } t > 0 \\ 0 & \text{otherwise} \end{cases} .$$

This generalizes the linear case and produces valid intrinsic (albeit not weakly) stationary semivariograms provided $0 \leq \lambda < 2$.

9. *Matérn* : The variogram for the Matérn class is given by

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2\left[1 - \frac{(2\sqrt{\nu}t\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(2\sqrt{\nu}t\phi)\right] & \text{if } t > 0 \\ \tau^2 & \text{otherwise} \end{cases} . \tag{2.8}$$

This class was originally suggested by Matérn (1960, 1986). Interest in it was revived by Handcock and Stein (1993) and Handcock and Wallis (1994),

model	Variogram, $\gamma(t)$
Linear	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 t & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Spherical	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t \geq 1/\phi \\ \tau^2 + \sigma^2 \left[\frac{3}{2}\phi t - \frac{1}{2}(\phi t)^3\right] & \text{if } 0 < t \leq 1/\phi \\ 0 & \text{otherwise} \end{cases}$
Exponential	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi t)) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Powered exponential	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\lvert\phi t\rvert^p)) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Gaussian	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi^2 t^2)) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Rational quadratic	$\gamma(t) = \begin{cases} \tau^2 + \frac{\sigma^2 t^2}{(1+\phi t^2)} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Wave	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \frac{\sin(\phi t)}{\phi t}) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Power law	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 t^\lambda & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Matérn	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 \left[1 - \frac{(2\sqrt{\nu}t\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(2\sqrt{\nu}t\phi)\right] & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Matérn at $\nu = 3/2$	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 \left[1 - (1 + \phi t) \exp(-\phi t)\right] & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$

Table 2.2 *Summary of variograms for common parametric isotropic models.*

who demonstrated attractive interpretations for ν as well as ϕ. Here $\nu > 0$ is a parameter controlling the smoothness of the realized random field (see Subsection 2.2.3) while ϕ is a spatial scale parameter. The function $\Gamma(\cdot)$ is the usual gamma function while K_ν is the modified Bessel function of order ν (see, e.g., Abramowitz and Stegun, 1965, Chapter 9). Implementations of this function are available in several C/C++ libraries and also in the R package geoR. Note that special cases of the above are the exponential ($\nu = 1/2$) and the Gaussian ($\nu \to \infty$). At $\nu = 3/2$ we obtain a closed form as well, namely $\gamma(t) = \tau^2 + \sigma^2 [1 - (1 + \phi t) \exp(-\phi t)]$ for $t > 0$, and τ^2 otherwise.

The covariance functions and variograms we have described in this subsection are conveniently summarized in Tables 2.1 and 2.2, respectively.

2.1.4 *Variogram model fitting*

Having seen a fairly large selection of models for the variogram, one might well wonder how we choose one of them for a given data set, or whether the data can really distinguish them (see Subsection 5.1.3 in this latter regard). Historically, a variogram model is chosen by plotting the *empirical semivariogram* (Matheron, 1963), a simple nonparametric estimate of the semivariogram, and then comparing it to the various theoretical shapes available from the choices in the previous subsection. The customary empirical semivariogram is

$$\widehat{\gamma}(t) = \frac{1}{2N(t)} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(t)} [Y(\mathbf{s}_i) - Y(\mathbf{s}_j)]^2 \; , \qquad (2.9)$$

where $N(t)$ is the set of pairs of points such that $||\mathbf{s}_i - \mathbf{s}_j|| = t$, and $|N(t)|$ is the number of pairs in this set. Notice that, unless the observations fall on a regular grid, the distances between the pairs will all be different, so this will not be a useful estimate as it stands. Instead we would "grid up" the t-space into intervals $I_1 = (0, t_1), I_2 = (t_1, t_2)$, and so forth, up to $I_K = (t_{K-1}, t_K)$ for some (possibly regular) grid $0 < t_1 < \cdots < t_K$. Representing the t values in each interval by its midpoint, we then alter our definition of $N(t)$ to

$$N(t_k) = \{(\mathbf{s}_i, \mathbf{s}_j) : ||\mathbf{s}_i - \mathbf{s}_j|| \in I_k\} \; , \; k = 1, \ldots, K \; .$$

Selection of an appropriate number of intervals K and location of the upper endpoint t_K is reminiscent of similar issues in histogram construction. Journel and Huijbregts (1979) recommend bins wide enough to capture at least 30 pairs per bin.

Clearly (2.9) is nothing but a method of moments (MOM) estimate, the semivariogram analogue of the usual sample variance estimate s^2. While very natural, there is reason to doubt that this is the best estimate of the semivariogram. Certainly it will be sensitive to outliers, and the sample average of the squared differences may be rather badly behaved since under a Gaussian distributional assumption for the $Y(\mathbf{s}_i)$, the squared differences will have a distribution that is a scale multiple of the heavily skewed χ^2_1 distribution. In this regard, Cressie and Hawkins (1980) proposed a robustified estimate that uses sample averages of $|Y(\mathbf{s}_i) - Y(\mathbf{s}_j)|^{1/2}$; this estimate is available in several software packages (see Section 2.5.1 below). Perhaps more uncomfortable is that (2.9) uses data differences, rather than the data itself. Also of concern is the fact that the components of the sum in (2.9) will be dependent within and across bins, and that $N(t_k)$ will vary across bins.

In any case, an empirical semivariogram estimate can be plotted, viewed, and an appropriately shaped theoretical variogram model can be fit to this "data." Since any empirical estimate naturally carries with it a significant amount of noise in addition to its signal, this fitting of a theoretical model has traditionally been as much art as science: in any given real data setting, any number of different models (exponential, Gaussian, spherical, etc.) may seem equally appropriate. Indeed, fitting has historically been done "by eye," or at best by using trial and error to choose values of nugget, sill, and range parameters that provide a good match to the empirical semivariogram (where the "goodness" can be judged visually or by using some least squares or similar criterion); again see Section 2.5.1. More formally, we could treat this as a statistical estimation problem, and use nonlinear maximization routines to find nugget, sill, and range parameters that minimize some goodness-of-fit criterion.

If we also have a distributional model for the data, we could use maximum likelihood (or restricted maximum likelihood, REML) to obtain sensible parameter estimates; see, e.g., Smith (2001) for details in the case of Gaussian data modeled with the various parametric variogram families outlined in Subsection 2.1.3. In Chapter 4 and Chapter 5 we shall see that the hierarchical Bayesian approach is broadly similar to this latter method, although it will often be easier and more intuitive to work directly with the covariance model $C(t)$, rather than changing to a partial likelihood in order to introduce the semivariogram.

2.2 Spatial process models $\star$

2.2.1 Formal modeling theory for spatial processes

When we write the collection of random variables $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ for some region of interest D or more generally $\{Y(\mathbf{s}) : \mathbf{s} \in \Re^r\}$, it is evident that we are envisioning a stochastic process indexed by $\mathbf{s}$. To capture spatial association it is also evident that these variables will be pairwise dependent with strength of dependence that is specified by their locations.

So, in fact, we have to determine the joint distribution for an uncountable number of random variables. In fact, we do this through specification of arbitrary finite dimensional distributions, i.e., for an arbitrary number of and choice of locations. Consistency of such specifications in terms of ensuring a unique joint distribution will rarely hold and will be difficult to establish. We avoid such technical concerns here by confining ourselves to Gaussian processes or to mixtures of such processes. In this case, all that is required is a valid correlation function, as we discuss below.

Again, to clarify the inference setting, in practice we will only observe $Y(\mathbf{s})$ at a finite set of locations, $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$. Based upon $\{Y(\mathbf{s}_i), i = 1, \ldots, n\}$, we seek to infer about the mean, variability, and association structure of the process. We also seek to predict $Y(\mathbf{s})$ at arbitrary unobserved locations. Since our focus is on hierarchical modeling, often the spatial process is introduced through random effects at the second stage of the modeling specification. In this case, we still have the same inferential questions but now the process is never actually observed. It is latent and the data, modeled at the first stage, helps us to learn about the process.

In this sense, we can make intuitive connections with familiar dynamic models (e.g., West and Harrison, 1997) where there is a latent state space model that is temporally updated. In fact, this reminds us of a critical difference between the one-dimensional time domain and the two-dimensional spatial domain: we have full order in the former, but only partial order in two or more dimensions.

The implications of this remark are substantial. Large sample analysis for time series usually lets time go to ∞. Asymptotics envision an increasing time domain. By contrast, large sample analysis for spatial process data usually envisions a fixed region with more and more points filling in this domain (so-called infill asymptotics). When applying increasing domain asymptotic results, we can assume that, as we collect more and more data, we can learn about temporal association at increasing distance in time. When applying infill asymptotic results for a fixed domain we can learn more and more about association as distance between points tends to 0. However, with a maximum distance fixed by the domain we cannot learn about association (in terms of consistent inference) at increasing distance. The former remark indicates that we may be able to do an increasingly better job with regard to spatial prediction at a given location. However, we need not be doing better in terms of inferring about other features of the process. See the work of Stein (1999a, 1999b) for a full technical discussion regarding such asymptotic results. Here, we view such concerns as providing encouragement for using a Bayesian framework for inference, since then we need not rely on any asymptotic theory for inference, but rather obtain exact inference given whatever data we have observed.

Before we turn to some technical discussion regarding covariance and correlation functions, we note that the above restriction to Gaussian processes enables several advantages. First, it allows very convenient distribution theory. Joint marginal and conditional distributions are all immediately obtained from standard theory once the mean and covariance structure have been specified. In fact, this is all we need to specify in order to determine all distributions. Also, as we shall see, in the context of hierarchical modeling, a Gaussian process assumption for spatial random effects introduced at the second stage of the model is very natural in the same way that independent random effects with variance components are customarily introduced in linear or generalized linear mixed models. From a technical point of view, as noted in Subsection 2.1.1, if we work with Gaussian processes and stationary models, strong stationarity is equivalent to weak stationarity. We will clarify these notions in the next subsection. Lastly, in most applications, it is difficult to criticize a Gaussian assumption. To argue this as simply as possible, in the absence of replication we have $\mathbf{Y} = (Y(s_1, \ldots, Y(s_n))$, a single realization from an n-dimensional distribution. With a sample size of one, how can we criticize *any* multivariate distributional specification (Gaussian or otherwise)?

Strictly speaking this last assertion is not quite true with a Gaussian process model. That is, the joint distribution is a multivariate normal with mean say $\mathbf{0}$, and a covariance matrix that is a parametric function of the parameters in the covariance function. When n is large enough, the effective sample size will also be large. We can obtain by linear transformation a set of approximately uncorrelated variables through which the adequacy of the normal assumption can be studied. We omit details.

2.2.2 *Covariance functions and spectra*

In order to specify a stationary process we must provide a valid covariance function. Here "valid" means that $c(\mathbf{h}) \equiv cov(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h}))$ is such that for any finite set of sites $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and for any $a_1, \ldots, a_n$,

$$Var\left[\sum_i a_i Y(s_i)\right] = \sum_{i,j} a_i a_j Cov(Y(\mathbf{s}_i), Y(\mathbf{s}_j)) = \sum_{i,j} a_i a_j c(\mathbf{s}_i - \mathbf{s}_j) \geq 0\,,$$

with strict inequality if not all the a_i are 0. That is, we need $c(\mathbf{h})$ to be a positive definite function.

Verifying the positive definiteness condition is evidently not routine. Fortunately, we have *Bochner's Theorem* (see, e.g., Gikhman and Skorokhod, 1974, p. 208), which provides a necessary and sufficient condition for $c(\mathbf{h})$ to be positive definite. This theorem is applicable for $\mathbf{h}$ in arbitrary r-dimensional Euclidean space, although our primary interest is in $r = 2$.

In general, for real-valued processes, Bochner's Theorem states that $c(\mathbf{h})$ is positive definite if and only if

$$c(\mathbf{h}) = \int \cos(\mathbf{w}^T\mathbf{h})\, G(d\mathbf{w})\,, \tag{2.10}$$

where G is a bounded, positive, symmetric about 0 measure in $\Re^r$. Then $c(\mathbf{0}) = \int Gd(\mathbf{w})$ becomes a normalizing constant, and $G(d\mathbf{w})/c(\mathbf{0})$ is referred to as the *spectral distribution* that induces $c(\mathbf{h})$. If $G(d\mathbf{w})$ has a density with respect to Lebesgue measure, i.e., $G(d\mathbf{w}) = g(\mathbf{w})d\mathbf{w}$, then $g(\mathbf{w})/c(\mathbf{0})$ is referred to as the *spectral density*. Evidently, (2.10) can be used to generate valid covariance functions; see (2.12) below. Of course, the behavioral implications associated with c arising from a given G will only be clear in special cases, and (2.10) will be integrable in closed form only in cases that are even more special.

Since $e^{i\mathbf{w}^T\mathbf{h}} = \cos(\mathbf{w}^T\mathbf{h}) + i\sin(\mathbf{w}^T\mathbf{h})$, we have $c(\mathbf{h}) = \int e^{i\mathbf{w}^T\mathbf{h}} G(d\mathbf{w})$. That is, the imaginary term disappears due to the symmetry of G around 0. In other words, $c(\mathbf{h})$ is a valid covariance function if and only if it is

the characteristic function of an r-dimensional symmetric random variable (random variable with a symmetric distribution). We note that if G is not assumed to be symmetric about $\mathbf{0}$, $c(\mathbf{h}) = \int e^{i\mathbf{w}^T\mathbf{h}} G(d\mathbf{w})$ still provides a valid covariance function (i.e., positive definite) but now for a complex-valued random process on $\Re^r$.

The Fourier transform of $c(\mathbf{h})$ is

$$\widehat{c}(\mathbf{w}) = \int e^{-i\mathbf{w}^T\mathbf{h}}\, c(\mathbf{h})d\mathbf{h}\,. \tag{2.11}$$

Applying the inversion formula, $c(\mathbf{h}) = (2\pi)^{-2} \int e^{i\mathbf{w}^T\mathbf{h}} \widehat{c}(\mathbf{w})d\mathbf{w}$, we see that $(2\pi)^{-r}\widehat{c}(\mathbf{w})/c(0) = g(\mathbf{w})$, the spectral density. Explicit computation of (2.11) is usually not possible except in special cases. However, approximate calculation is available through the fast Fourier transform (FFT); see Appendix A, Section A.4. Expression (2.11) can be used to check whether a given $c(\mathbf{h})$ is valid: we simply compute $\widehat{c}(\mathbf{w})$ and check whether it is positive and integrable (so it is indeed a density up to normalization).

The one-to-one relationship between $c(\mathbf{h})$ and $g(\mathbf{w})$ enables examination of spatial processes in the spectral domain rather than in the observational domain. Computation of $g(\mathbf{w})$ can often be expedited through fast Fourier transforms; g can be estimated using the so-called *periodogram*. Likelihoods can be obtained approximately in the spectral domain enabling inference to be carried out in this domain. See, e.g., Guyon (1995) or Stein (1999a) for a full development. Likelihood evaluation is much faster in the spectral domain. However, in this book we confine ourselves to the observational domain because of concerns regarding the accuracy associated with approximation in the spectral domain (e.g., the likelihood of Whittle, 1954), and with the ad hoc creation of the periodogram (e.g., how many low frequencies are ignored). We do however note that the spectral domain may afford the best potential for handling computation associated with large data sets.

Isotropic covariance functions, i.e., $c(||\mathbf{h}||)$, where $||\mathbf{h}||$ denotes the length of $\mathbf{h}$, are the most frequently adopted choice within the stationary class. There are various direct methods for checking the permissibility of isotropic covariance and variogram specifications. See, e.g., Armstrong and Diamond (1984), Christakos (1984), and McBratney and Webster (1986). Again denoting $||\mathbf{h}||$ by t for notational simplicity, recall that Tables 2.1 and 2.2 provide the covariance function $C(t)$ and variogram $\gamma(t)$, respectively, for the widely encountered parametric istropic choices that were initially presented in Subsection 2.1.3.

It is noteworthy that an isotropic covariance function that is valid in dimension r need not be valid in dimension $r+1$. The intuition may be gleaned by considering $r = 1$ versus $r = 2$. For three points, in one-dimensional space, given the distances separating points 1 and 2 (d_{12}) and points 2 and 3 (d_{23}), then the distance separating points 1 and 3 d_{13} is either $d_{12} + d_{23}$ or $|d_{12} - d_{23}|$. But in two-dimensional space, given d_{12} and d_{23}, d_{13} can take any value in $\Re^+$ (subject to triangle inequality). With increasing dimension more sets of interlocation distances are possible for a given number of locations; it will be more difficult for a function to satisfy the positive definiteness condition. Armstrong and Jabin (1981) provide an explicit example that we defer to Exercise 3.

There are isotropic correlation functions that are valid in all dimensions. The Gaussian correlation function, $k(\|h\|) = \exp(-\phi\,\|h\|^2)$ is an example. It is the characteristic function associated with r i.i.d. normal random variables, each with variance $1/(2\phi)$ for any r. More generally, the powered exponential, $\exp(-\phi\,\|h\|^\alpha)$, $0 < \alpha \leq 2$ (and hence the exponential correlation function) is valid for any r. The Cauchy correlation function is also valid in any dimension.

Rather than seeking isotropic correlation functions that are valid in all dimensions, we might seek all valid isotropic correlation function in a particular dimension r. Matérn (1960, 1986) provides the general result. The set of $c(\|h\|)$ of the form

$$c(\|h\|) = \int_0^\infty \left(\frac{2}{w\,\|\mathbf{h}\|}\right)^\alpha \Gamma(\nu+1) J_\nu(w\,\|\mathbf{h}\|) G(dw)\,, \tag{2.12}$$

where G is nondecreasing and integrable on $\Re^+$, J_ν is the Bessel function of the first kind of order ν, and $\nu = (r-2)/2$ provides all valid isotropic correlation functions on $\Re^r$.

When $r = 2$, $v = 0$ so that arbitrary correlation functions in two-dimensional space arise as scale mixtures of Bessel functions of order 0. In particular, $J_0(d) = \sum_{k=0}^\infty \frac{(-1)^k}{(k!)^2}\left(\frac{d}{2}\right)^{k/2}$. J_0 decreases from 1 at $d = 0$ and will oscillate above and below 0 with amplitudes and frequencies that are diminishing as d increases (see Figure 5.1 in Section 5.1). Typically, correlation functions that are monotonic and decreasing to 0 are chosen but, apparently, valid correlation functions can permit negative associations with w determining the scale in distance space. Such behavior might be appropriate in certain applications.

The form in (2.12) at $\nu = 0$ was exploited in Shapiro and Botha (1981) and Ver Hoef and Barry (1998) to develop "nonparametric" variogram models and "black box" kriging. It was employed in Ecker and Gelfand (1997) to obtain flexible spatial process models within which to do inference from a Bayesian perspective (see Subsection 5.1.3).

If we confine ourselves to strictly monotonic isotropic covariance functions then we can introduce the notion of a range. As described above, the range is conceptualized as the distance beyond which association becomes negligible. If the covariance function reaches 0 in a finite distance, then we refer to this distance as the range. However, as Table 2.1 reveals, we customarily work with covariance functions that attain 0 asymptotically as $\|\mathbf{h}\| \to \infty$. In this case, it is common to define the range as the distance beyond which correlation is less than .05, and this is the definition we employ in the sequel. So if ρ is the correlation function, then writing the range as R we solve $\rho(R;\boldsymbol{\theta}) = .05$, where $\boldsymbol{\theta}$ denotes the parameters in the correlation function. Therefore, R is an implicit function of the parameter $\boldsymbol{\theta}$.

We do note that some authors define the range through the variogram, i.e., the distance at which the variogram reaches .95 of its sill. That is, we would solve $\gamma(R) = .95(\sigma^2 + \tau^2)$. Note, however, that if we rewrite this equation in terms of the correlation function we obtain $\tau^2 + \sigma^2(1 - \rho(R;\boldsymbol{\theta})) = .95(\tau^2 + \sigma^2)$, so that $\rho(R;\boldsymbol{\theta}) = .05\left(\frac{\sigma^2+\tau^2}{\sigma^2}\right)$. Evidently, the solution to this equation is quite different from the solution to the above equation. In fact, this latter equation may not be solvable, e.g., if $\sigma^2/(\sigma^2 + \tau^2) \leq .05$, the case of very weak "spatial story" in the model. As such, one might argue that a spatial model is inappropriate in this case. However, with σ^2 and τ^2 unknown, it seems safer to work with the former definition.

We note that one can offer constructive strategies to build larger classes of correlation functions. Three approaches are mixing, products, and convolution. Mixing notes simply that if $C_1, \ldots, C_m$ are valid correlation functions in $\Re^r$ and if $\sum_{i=1}^m p_i = 1$, $p_i > 0$, then $C(\mathbf{h}) = \sum_{i=1}^m p_i C_i(\mathbf{h})$ is also a valid correlation function in $\Re^r$. This follows since $C(\mathbf{h})$ is the characteristic function associated with $\sum p_i f_i(\mathbf{x})$, where $f_i(\mathbf{x})$ is the symmetric about 0 density in r-dimensional space associated with $C_i(\mathbf{h})$.

Using products simply notes that again if $c_1, \ldots, c_n$ are valid in $\Re^r$, then $\prod_{i=1}^m c_i$ is a valid correlation function in $\Re^r$. This follows since $\prod_{i=1}^m c_i(n)$ is the characteristic function associated with $V = \sum_{i=1}^m V_i$ where the V_i are independent with V_i having characteristic function $c_i(\mathbf{h})$.

Convolution simply recognizes that if c_1 and c_2 are valid correlation functions in $\Re^r$, then $c_{12}(\mathbf{h}) = \int c_1(\mathbf{h} - \mathbf{t})c_2(\mathbf{t})d\mathbf{t}$ is a valid correlation function in $\Re^r$. The argument here is to look at the Laplace transform of $c_{12}(\mathbf{h})$. That is,

$$\begin{aligned} \widehat{c}_{12}(\mathbf{w}) &= \textstyle\int e^{-i\mathbf{w}^T\mathbf{h}} c_{12}(\mathbf{h})d\mathbf{h} \\ &= \textstyle\int e^{-i\mathbf{w}^T\mathbf{h}} \int c_1(\mathbf{h}-\mathbf{t})c_2(\mathbf{t})d\mathbf{t}d\mathbf{h} \\ &= \widehat{c}_1(\mathbf{w}) \cdot \widehat{c}_2(\mathbf{w})\,, \end{aligned}$$

where $\widehat{c}_i(\mathbf{w})$ is the Laplace transform of $c_i(\mathbf{h})$ for $i = 1, 2$. But then $c_{12}(\mathbf{h}) = (2\pi)^{-2}\int e^{i\mathbf{w}^T\mathbf{h}}\widehat{c}_1(\mathbf{w})\widehat{c}_2(\mathbf{w})d\mathbf{w}$. Now $\widehat{c}_1(\mathbf{w})$ and $\widehat{c}_2(\mathbf{w})$ are both symmetric about $\mathbf{0}$ since, up to a constant, they are the spectral densities associated with $c_1(\mathbf{h})$ and $c_2(\mathbf{h})$, respectively. Hence, $c_{12}(\mathbf{h}) = \int \cos \mathbf{w}^T\mathbf{h} G(d\mathbf{w})$ where $G(d\mathbf{w}) = (2\pi)^{-2}\widehat{c}_1(\mathbf{w})^2\widehat{c}_2(\mathbf{w})d\mathbf{w}$.

Thus, from (2.10), $c_{12}(\mathbf{h})$ is a valid correlation function, i.e., G is a bounded, positive, symmetric about 0 measure on $\Re^2$. In fact, if c_1 and c_2 are isotropic then c_{12} is as well; we leave this verification as Exercise 5.

2.2.3 Smoothness of process realizations

How does one select among the various choices of correlation functions? Usual model selection criteria will typically find it difficult to distinguish, say, among one-parameter isotropic scale choices such as the exponential, Gaussian, or Cauchy. Ecker and Gelfand (1997) provide some graphical illustration showing that, through suitable alignment of parameters, the correlation curves will be very close to each other. Of course, in comparing choices with parametrizations of differing dimensions (e.g., correlation functions developed using results from the previous section), we will need to employ a selection criterion that penalizes complexity and rewards parsimony (see Section 4.2.3).

An alternative perspective is to make the selection based upon theoretical considerations. This possibility arises from the powerful fact that the choice of correlation function determines the smoothness of realizations from the spatial process. More precisely, a process realization is viewed as a random surface over the region. By choice of c we can ensure that these realizations will be almost surely continuous, or mean square continuous, or mean square differentiable, and so on. Of course, at best the process is only observed at finitely many locations. (At worst, it is never observed, e.g., when the spatial process is used to model random spatial effects.) So, it is not possible to "see" the smoothness of the process realization. Elegant theory, developed in Kent (1989), Stein (1999a) and extended in Banerjee and Gelfand (2003), clarifies the relationship between the choice of correlation function and such smoothness. We provide a bit of this theory below, with further discussion in Section 10.1. For now, the key point is that, according to the process being modeled, we may, for instance, anticipate surfaces not be continuous (as with digital elevation models in the presence of gorges, escarpments, or other topographic features), or to be differentiable (as in studying land value gradients or temperature gradients). We can choose a correlation function to essentially ensure such behavior.

Of particular interest in this regard is the Matérn class of covariance functions. The parameter v (see Table 2.1) is, in fact, a smoothness parameter. In two-dimensional space, the greatest integer in v indicates the number of times process realizations will be mean square differentiable. In particular, since $v = \infty$ corresponds to the Gaussian correlation function, the implication is that use of the Gaussian correlation function results in process realizations that are mean square analytic, which may be too smooth to be appropriate in practice. That is, it is possible to predict $Y(\mathbf{s})$ perfectly for all $\mathbf{s} \in \Re^2$ based upon observing $Y(\mathbf{s})$ in an arbitrarily small neighborhood. Expressed in a different way, use of the Matérn covariance function as a model enables the data to inform about v; we can learn about process smoothness despite observing the process at only a finite number of locations.

Hence, we follow Stein (1999a) in recommending the Matérn class as a general tool for building spatial models. The computation of this function requires evaluation of a modified Bessel function. In fact, evaluation will be done repeatedly to obtain a covariance matrix associated with n locations, and then iteratively if a model is fit via MCMC methods. This may appear off-putting but, in fact, such computation can be done efficiently using expansions to approximate $K_v(\cdot)$ (Abramowitz and Stegun, p. 435), or working through the inversion formula below (2.11), which in this case becomes

$$2\left(\frac{\phi\,\|\mathbf{h}\|}{2}\right)^{\nu} \frac{K_v(\phi(\|\mathbf{h}\|))}{\phi^{2\nu}\Gamma(v+\frac{r}{2})} = \int_{\Re^r} e^{i\mathbf{w}^T\mathbf{h}}(\phi^2 + \|\mathbf{w}\|^2)^{-(v+r/2)}d\mathbf{w}\ , \quad (2.13)$$

where K_v is the modified Bessel function of order ν.

Computation of (2.13) is discussed further in Appendix Section A.4. In particular, the right side of (2.13) is readily approximated using fast Fourier transforms. Again, we revisit process smoothness in Section 10.1.

2.2.4 Directional derivative processes

The previous section offered discussion intended to clarify, for a spatial process, the connection between correlation function and smoothness of process realizations. When realizations are mean square differentiable, we can think about a directional derivative process. That is, for a given direction, at each location we can define a random variable that is the directional derivative of the original process at that location in the given direction. The entire collection of random variables can again be shown to be a spatial process. We offer brief development below but note that, intuitively, such variables would be created through limits of finite differences. In other words, we can also formalize a finite difference process in a given direction. The value of formalizing such processes lies in the possibility of assessing where, in a region of interest, there are sharp gradients and in which directions. They also enable us to work at different scales of resolution. Application could involve land-value gradients away from a central business district, temperature gradients in a north-south direction as mentioned above, or perhaps the maximum gradient at a location and the direction of that gradient, in order to identify zones of rapid change (boundary analysis). Some detail in the development of directional derivative processes appears in Subsection 10.1.2.

2.2.5 Anisotropy

Geometric anisotropy

Stationary correlation functions extend the class of correlation functions from isotropy where association only depends upon distance to association that depends upon the separation vector between locations. As a result, association depends upon direction. An illustrative example is the class of geometric anisotropic correlation functions where we set

$$c(\mathbf{s}-\mathbf{s}') = \sigma^2\rho((\mathbf{s}-\mathbf{s}')^T B(\mathbf{s}-\mathbf{s}'))\ . \tag{2.14}$$

In (2.14), B is positive definite with ρ a valid correlation function in $\Re^r$ (say, from Table 2.1). We would omit the range/decay parameter since it can be incorporated into B. When $r = 2$ we obtain a specification with three parameters rather than one. Contours of constant association arising from c in (2.14) are elliptical. In particular, the contour corresponding to $\rho = .05$ provides the range in each spatial direction. Ecker and Gelfand (1997) provide the details for Bayesian modeling and inference incorporating (2.14); see also Subsection 5.1.4.

Following the discussion in Subsection 2.2.2, we can extend geometric anisotropy to *product* geometric anisotropy. In the simplest case, we would set

$$c(s-s') = \sigma^2\ \rho_1((\mathbf{s}-\mathbf{s}')^T B_1(\mathbf{s}-\mathbf{s}'))\ \rho_2((\mathbf{s}-\mathbf{s}')^T B_2(\mathbf{s}-\mathbf{s}'))\ ,$$

noting that c is valid since it arises as a product of valid covariance functions. See Ecker and Gelfand (2003) for further details and examples.

Other notions of anisotropy

In a more general discussion, Zimmerman (1993) suggests three different notions of anisotropy: *sill* anisotropy, *nugget* anisotropy, and *range* anisotropy. More precisely, working with a variogram $\gamma(\mathbf{h})$, let $\mathbf{h}$ be an arbitrary separation vector so that $\mathbf{h}/\,\|\mathbf{h}\|$ is a unit vector in $\mathbf{h}$'s direction. Consider $\gamma(c\mathbf{h}/\,\|\mathbf{h}\|)$. Let $c \to \infty$ and suppose $\lim_{c\to\infty}\gamma(c\mathbf{h}/\,\|\mathbf{h}\|)$ depends upon $\mathbf{h}$. This situation is naturally referred to a sill anisotropy. If we work with the usual relationship $\gamma(c\mathbf{h}/\,\|\mathbf{h}\|) = \tau^2+\sigma^2\left(1-\rho\left(c\frac{\mathbf{h}}{\|\mathbf{h}\|}\right)\right)$, then, in some directions, ρ must not go to 0 as $c \to \infty$. If this can be the case, then ergodicity assumptions (i.e., convergence assumptions associated with averaging) will be violated. If this can be the case, then perhaps the constant mean assumption, implicit for the variogram, does not hold. Alternatively, it is also possible that the constant nugget assumption fails.

Instead, let $c \to 0$ and suppose $\lim_{c\to 0}\gamma(c\mathbf{h}/\,\|\mathbf{h}\|)$ depends upon $\mathbf{h}$. This situation is referred to as nugget anisotropy. Since, by definition, ρ must go to 1 as $c \to 0$. This says that the measurement errors that are assumed uncorrelated with common variance may be correlated. More generally, a simple white noise process model for the nonspatial errors is not appropriate.

A third type of anisotropy is range anisotropy where the range depends upon direction. Zimmerman (1993) asserts that "this is the form most often seen in practice." Geometric anisotropy and the more general product geometric anisotropy from the previous subsections are illustrative cases. However, given the various constructive strategies offered in Subsection 2.2.2 to create more general stationary covariance functions, we can envision nongeometric range anisotropy, implying general correlation function or variogram contours in $\Re^2$. However, due to the positive definiteness restriction on the correlation function, the extent of possible contour shapes is still rather limited.

Lastly, motivated by directional variograms (see Subsection 2.3.2), some authors propose the idea of nested models (see Zimmerman, 1993, and the references therein). That is, for each separation vector there is an associated angle with, say, the x-axis, which by symmetry considerations can be restricted to $[0, \pi)$. Partitioning this interval into a set of angle classes, a different variogram model is assumed to operate for each class. In terms of correlations, this would imply a different covariance function is operating for each angle class. But evidently this does not define a valid process model: the resulting covariance matrix for an arbitrary set of locations need not be positive definite.

This can be seen with as few as three points and two angle classes. Let $(\mathbf{s}_1, \mathbf{s}_2)$ belong to one angle class with $(\mathbf{s}_1, \mathbf{s}_3)$ and $(\mathbf{s}_2, \mathbf{s}_3)$ in the other. With exponential isotropic correlation functions in each class by choosing ϕ_1 and ϕ_2 appropriately we can make $\rho(\mathbf{s}_1-\mathbf{s}_2) \approx 0$ while $\rho(\mathbf{s}_1-\mathbf{s}_3) = \rho(\mathbf{s}_2-\mathbf{s}_3) \approx 0.8$. A quick calculation shows that the resulting 3×3 covariance (correlation) matrix is not positive definite. So, in terms of being able to write proper joint distributions for the resulting data, nested models are inappropriate; they do not provide an extension of isotropy that allows for likelihood based inference.

2.3 Exploratory approaches for point-referenced data

2.3.1 Basic techniques

Exploratory data analysis (EDA) tools are routinely implemented in the process of analyzing one- and two-sample data sets, regression studies, generalized linear models, etc. (see, e.g., Chambers et al., 1983; Hoaglin, Mosteller, and Tukey, 1983, 1985; Aiktin et al., 1989). Similarly, such tools are appropriate for analyzing point-referenced spatial data.

For continuous data, the starting point is the so-called "first law of geostatistics." Figure 2.2 illustrates this "law" in a one-dimensional setting. The data is partitioned into a mean term and an error term. The mean corresponds to global (or *first-order*) behavior, while the error captures local (or *second-order*) behavior through a covariance function. EDA tools examine both first- and second-order behavior.

The law also clarifies that spatial association in the data, $Y(\mathbf{s})$, need

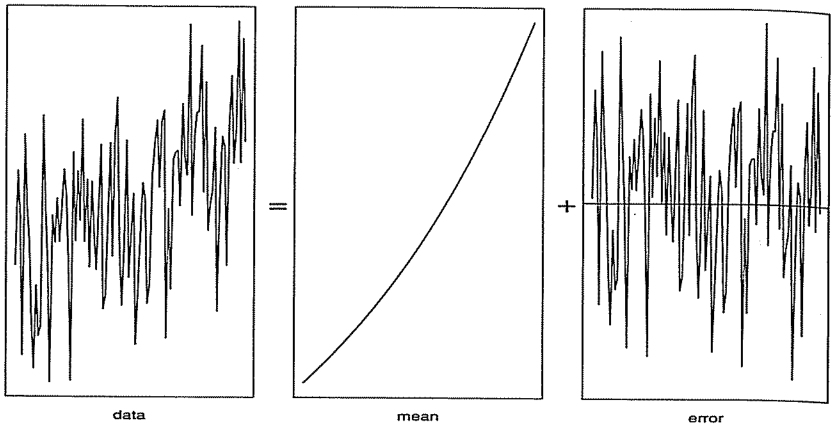


Figure 2.2 *Illustration of the first law of geostatistics.*

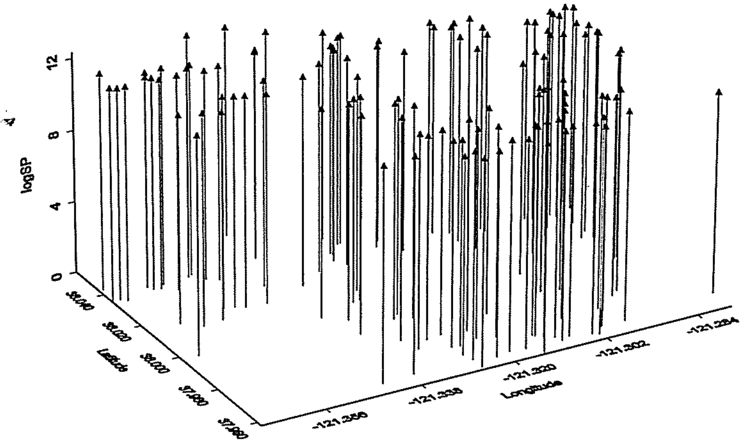


Figure 2.3 *Illustrative three-dimensional "drop line" scatterplot, scallop data.*

not resemble spatial association in the residuals, $\epsilon(\mathbf{s})$. That is, spatial association in the $Y(\mathbf{s})$ corresponds to looking at $E(Y(\mathbf{s}) - \mu)(Y(\mathbf{s}') - \mu)$, while spatial structure in the $\epsilon(\mathbf{s})$ corresponds to looking at $E(Y(\mathbf{s}) - \mu(\mathbf{s}))(Y(\mathbf{s}') - \mu(\mathbf{s}'))$. The difference between the former and the latter is $(\mu - \mu(\mathbf{s}))(\mu - \mu(\mathbf{s}'))$, which need not be negligible.

Certainly an initial exploratory display should be a simple map of the locations themselves. We need to assess how *regular* the arrangement of the points is. Next, some authors would recommend a stem-and-leaf display of the $Y(\mathbf{s})$. This plot is evidently nonspatial and is customarily for observations which are i.i.d. We expect both nonconstant mean and spatial dependence, but such a plot may at least suggest potential outliers. Next we might develop a three-dimensional "drop line" scatterplot of $Y(\mathbf{s}_i)$ versus $\mathbf{s}_i$, which we could convert to a three-dimensional surface plot or perhaps a contour plot as a *smoothed* summary. Examples of these three plots are shown for a sample of 120 log-transformed home selling prices in Stockton, CA, in Figures 2.3, 2.4, and 2.5, respectively. However, as the preceding paragraph clarifies, such displays may be deceiving. They may show spatial pattern that will disappear after $\mu(\mathbf{s})$ is fitted, or perhaps vice versa. It seems more sensible to study spatial pattern in the residuals.

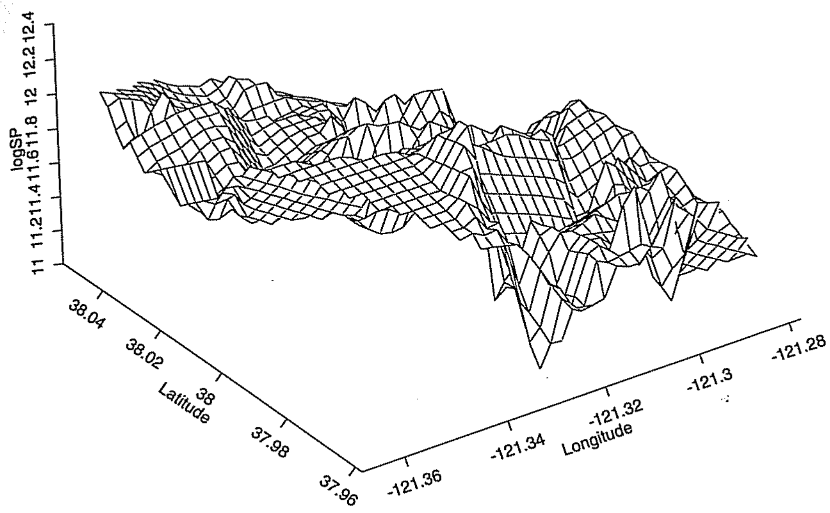


Figure 2.4 *Illustrative three-dimensional surface ("perspective") plot, Stockton real estate data.*

In exploring $\mu(\mathbf{s})$ we may have two types of information at location $\mathbf{s}$. One is the purely geographic information, i.e., the geocoded location expressed in latitude and longitude or as projected coordinates such as eastings and northings (Subsection 1.2.1 above). The other will be features relevant for explaining the $Y(\mathbf{s})$ at $\mathbf{s}$. For instance, if $Y(\mathbf{s})$ is a pollution concentration, then elevation, temperature, and wind information at $\mathbf{s}$ could well be useful and important. If instead $Y(\mathbf{s})$ is the selling price of a single-family home at $\mathbf{s}$, then characteristics of the home (square feet, age, number of bathrooms, etc.) would be useful.

When the mean is described purely through geographic information, $\mu(\mathbf{s})$ is referred to as a *trend surface*. When $\mathbf{s} \in \Re^2$, the surface is usually developed as a bivariate polynomial. For data that is roughly gridded (or can

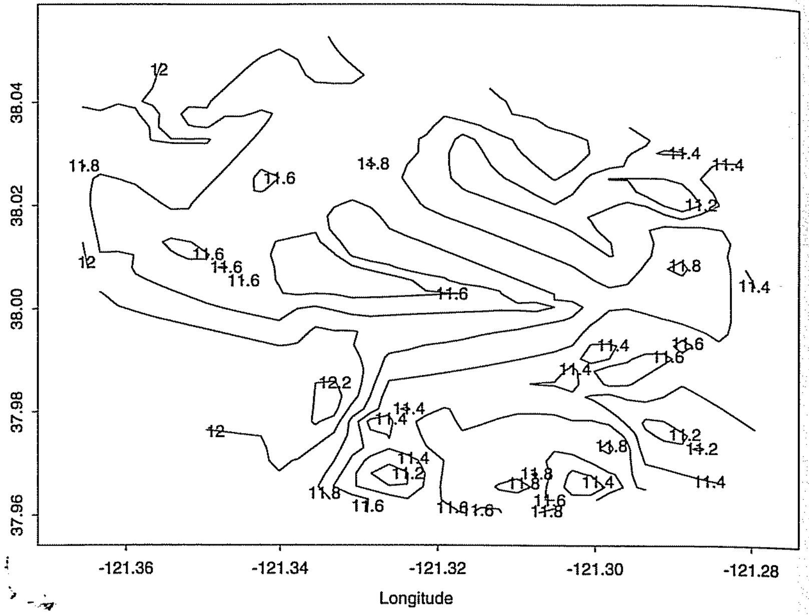


Figure 2.5 *Illustrative contour plot, Stockton real estate data.*

be assigned to row and column bins by overlaying a regular lattice on the points), we can make row and column boxplots looking for trend. Plotting these boxplots versus their center could clarify the existence and nature of such trend. In fact, median polishing (see, e.g., Hoaglin, Mosteller, and Tukey, 1985) could be used to extract row and column effects, and also to see if a multiplicative trend surface term is useful; see Cressie (1983, pp. 46–48) in this regard.

Figures 2.6 and 2.7 illustrate the row and column boxplot approach for a data set previously considered by Diggle and Ribeiro (2002). The response variable is the surface elevation ("height") at 52 locations on a regular grid within a 310-foot square (and where the mesh of the grid is 50 feet). The plots reveals some evidence of spatial pattern as we move along the rows, but not along the columns of the regular grid.

To assess small-scale behavior, some authors recommend creating the *semivariogram cloud*, i.e., a plot of $(Y(\mathbf{s}_i)-Y(\mathbf{s}_j))^2$ versus $||\mathbf{s}_i-\mathbf{s}_j||$. Usually this cloud is too "noisy" to reveal very much; see, e.g., Figure 5.2. The empirical semivariogram (2.9) is preferable in terms of reducing some of the noise, and can be a helpful tool in seeing the presence of spatial structure. Again, the caveat above suggests employing it for residuals (not the data itself) unless a constant mean is appropriate.

An empirical (nonparametric) covariance estimate, analogous to (2.9), is

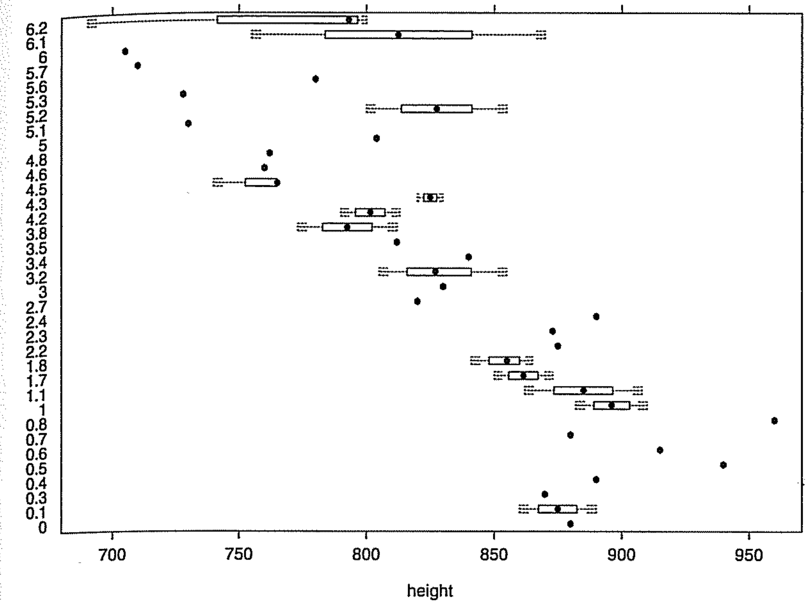


Figure 2.6 *Illustrative row box plots, Diggle and Ribeiro (2002) surface elevation data.*

also available. Creating bins as in this earlier approach, define

$$\widehat{c}(t_k) = \frac{1}{N_k} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(t_k)} (Y(\mathbf{s}_i)-\bar{Y})(Y(\mathbf{s}_j)-\bar{Y}) \, , \tag{2.15}$$

where again $N(t_k) = \{(\mathbf{s}_i,\mathbf{s}_j) : ||\mathbf{s}_i-\mathbf{s}_j|| \in I_k\}$ for $k = 1,\dots,K$, I_k indexes the kth bin, and there are N_k pairs of points falling in this bin. Equation (2.15) is a spatial generalization of a lagged autocorrelation in time series analysis. Since $\widehat{c}$ uses a common $\bar{Y}$ for all $Y(\mathbf{s}_i)$, it may be safer to employ (2.15) on the residuals. Two further issues arise: first, what should we define $\widehat{c}(0)$ to be, and second, regardless of this choice, the fact that $\widehat{\gamma}(t_k)$ does *not* equal $\widehat{c}(0)-\widehat{c}(t_k)$, $k = 1,\dots,K$. Details for both of these issues are left to Exercise 6.

Again, with a regular grid or binning we can create "same-lag" scatterplots. These are plots of $Y(\mathbf{s}_i+h\mathbf{e})$ versus $Y(\mathbf{s}_i)$ for a fixed h and a fixed unit vector $\mathbf{e}$. Comparisons among such plots may reveal the presence of anisotropy and perhaps nonstationarity.

Lastly, suppose we attach a neighborhood to each point. We can then compute the sample mean and variance for the points in the neighborhood, and even a sample correlation coefficient using all pairs of data in the neighborhood. Plots of each of them versus location can be informative.

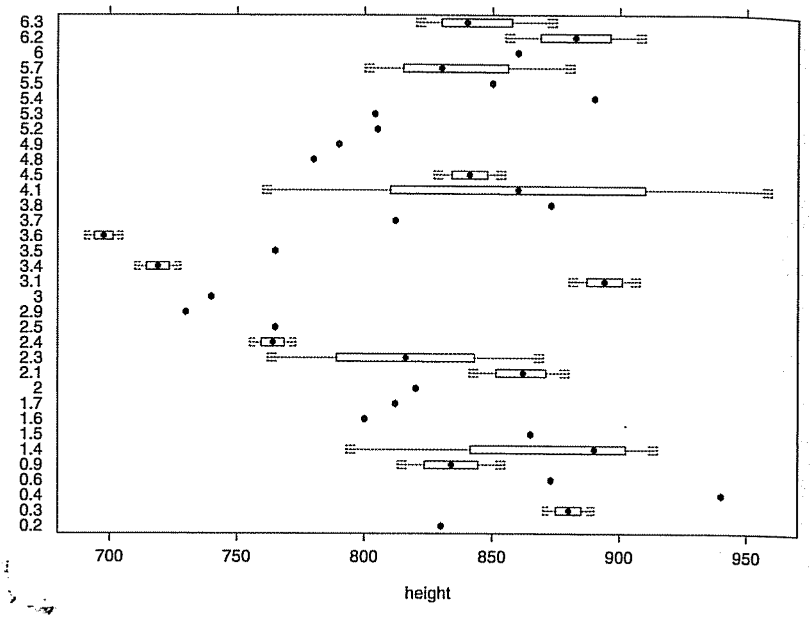


Figure 2.7 *Illustrative column box plots, Diggle and Ribeiro (2002) surface elevation data.*

The first may give some idea regarding how the mean structure changes across the study region. Plots of the second and third may provide evidence of nonstationarity. Implicit in extracting useful information from these plots is a roughly constant local mean. If $\mu(\mathbf{s})$ is to be a trend surface, this is plausible. But if $\mu(\mathbf{s})$ is a function of some geographic variables at $\mathbf{s}$ (say, home characteristics), then use of residuals would be preferable.

2.3.2 Assessing anisotropy

We illustrate various EDA techniques to assess anisotropy using sampling of scallop abundance on the continental shelf off the coastline of the northeastern U.S. The data from this survey, conducted by the Northeast Fisheries Science Center of the National Marine Fisheries Service, is available within the `S+SpatialStats` package; see Subsection 2.5.1. Figure 2.8 shows the sampling sites for 1990 and 1993.

Directional semivariograms and rose diagrams

The most common EDA technique for assessing anisotropy involves use of directional semivariograms. Typically, one chooses angle classes $\eta_i \pm \epsilon$, $i = 1, \ldots, L$ where ϵ is the halfwidth of the angle class and L is the number of angle classes. For example, a common choice of angle classes involves the four cardinal directions measured counterclockwise from the x-axis (0°, 45°, 90°, and 135°) where ϵ is 22.5°. Journel and Froidevaux (1982) display directional semivariograms at angles 35°, 60°, 125°, and 150° in deducing anistropy for a tungsten deposit. While knowledge of the underlying spatial characteristics of region D is invaluable in choosing directions, often the choice of the number of angle classes and the directions seems to be arbitrary.

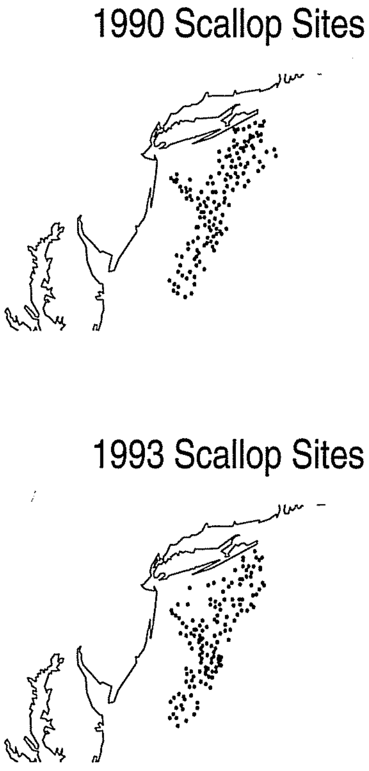


Figure 2.8 *Sites sampled in the Atlantic Ocean for 1990 and 1993 scallop catch data.*

For a given angle class, the Matheron empirical semivariogram (2.9) can be used to provide a directional semivariogram for angle η_i. Theoretically, all types of anisotropy can be assessed from these directional semivariograms; however, in practice determining whether the sill, nugget, and/or range varies with direction can be difficult. Figure 2.9(a) illustrates directional semivariograms for the 1990 scallop data in the four cardinal directions. Note that the semivariogram points are connected only to aid

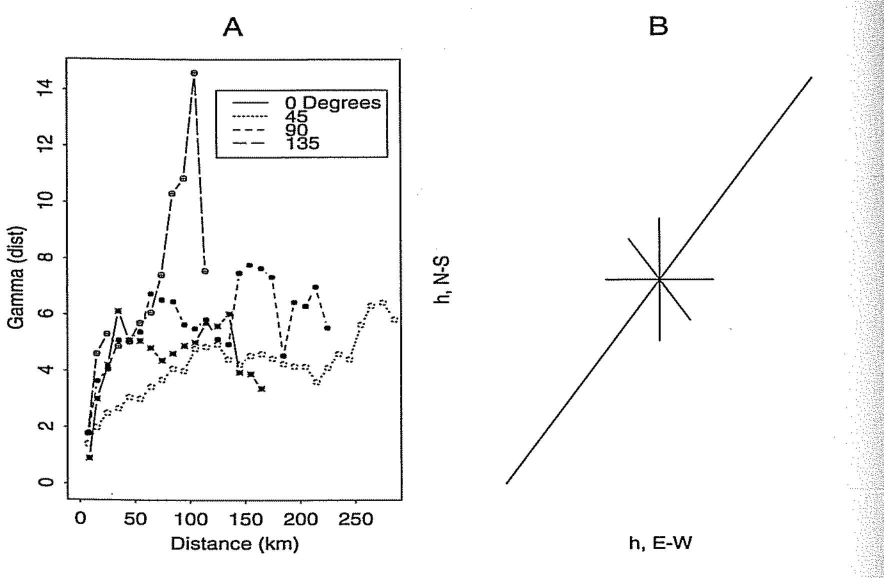


Figure 2.9 *Directional semivariograms (a) and a rose diagram (b) for the 1990 scallop data.*

comparison. Possible conclusions are: the variability in the 45° direction (parallel to the coastline) is significantly less than in the other three directions and the variability perpendicular to the coastline (135°) is very erratic, possibly exhibiting sill anisotropy. We caution however that it is dangerous to read too much significance and interpretation into directional variograms. No sample sizes (and thus no assessments of variability) are attached to these pictures. Directional variograms from data generated under a simple isotropic model will routinely exhibit differences of the magnitudes seen in Figure 2.9(a). Furthermore, it seems difficult to draw any conclusions regarding the presence of geometric anisotropy from this figure.

A rose diagram (Isaaks and Srivastava, 1989, pp. 151–154) can be created from the directional semivariograms to evaluate geometric anisotropy. At an arbitrarily selected γ^*, for a directional semivariogram at angle η, the distance d^* at which the directional semivariogram attains γ^* can be interpolated. Then, the rose diagram is a plot of angle η and corresponding distance d^* in polar cordinates. If an elliptical contour describes the extremities of the rose diagram reasonably well, then the process exhibits geometric anisotropy. For instance, the rose diagram for the 1990 scallop data is presented in Figure 2.9(b) using the γ^* contour of 4.5. It is approximately elliptical, oriented parallel to the coastline ($\approx 45°$) with a ratio of major to minor ellipse axes of about 4.

Empirical semivariogram contour (ESC) plots

A more informative method for assessing anisotropy is a contour plot of the empirical semivariogram surface in $\Re^2$. Such plots are mentioned informally in Isaaks and Srivastava (1989, pp. 149–151) and in Haining (1990, pp. 284–286); the former call them contour maps of the grouped variogram values, the latter an isarithmic plot of the semivariogram. Following Ecker and Gelfand (1999), we formalize such a plot here calling it an *empirical semivariogram contour* (ESC) plot. For each of the $\frac{N(N-1)}{2}$ pairs of sites in $\Re^2$, calculate h_x and h_y, the separation distances along each axis. Since the sign of h_y depends upon the arbitrary order in which the two sites are compared, we demand that $h_y \geq 0$. (We could alternatively demand that $h_x \geq 0$.) That is, we take $(-h_x, -h_y)$ when $h_y < 0$. These separation distances are then aggregated into rectangular bins B_{ij} where the empirical semivariogram values for the (i,j)th bin are calculated by

$$\gamma^*_{ij} = \frac{1}{2N_{B_{ij}}} \sum_{\{(k,l):(\mathbf{s}_k-\mathbf{s}_l)\in B_{ij}\}} \left(Y(\mathbf{s}_k) - Y(\mathbf{s}_l)\right)^2 , \qquad (2.16)$$

where $N_{B_{ij}}$ equals the number of sites in bin B_{ij}. Because we force $h_y \geq 0$ with h_x unrestricted, we make the bin width on the y-axis half of that for the x-axis. We also force the middle class on the x-axis to be centered around zero. Upon labeling the center of the (i,j)th bin by (x_i, y_j), a three dimensional plot of γ^*_{ij} versus (x_i, y_j) yields an empirical semivariogram surface. Smoothing this surface using, for example, the algorithm of Akima (1978) available in the S-plus software package (see Subsection 2.5.1) produces a contour plot that we call the ESC plot. A symmetrized version of the ESC plot can be created by reflecting the upper left quadrant to the lower right and the upper right quadrant to the lower left.

The ESC plot can be used to assess departures from isotropy; isotropy is depicted by circular contours while elliptical contours capture geometric anisotropy. A rose diagram traces only one arbitrarily selected contour of this plot. A possible drawback to the ESC plot is the occurrence of sparse counts in extreme bins. However, these bins may be trimmed before smoothing if desired. Concerned that use of geographic coordinates could introduce artificial anisotropy (since 1° latitude $\neq$ 1° longitude in the northeastern United States), we have employed a Universal Transverse Mercator (UTM) projection to kilometers in the E-W and N-S axes (see Subsection 1.2.1).

Figure 2.10 is the empirical semivariogram contour plot constructed using x-axis width of 30 kilometers for the 1993 scallop data. We have overlaid this contour plot on the bin centers with their respective counts. Note that using empirical semivariogram values in the row of the ESC plot, where $h_y \approx 0$ provides an alternative to the usual 0° directional semivariogram. The latter directional semivariograms are based on a polar representation

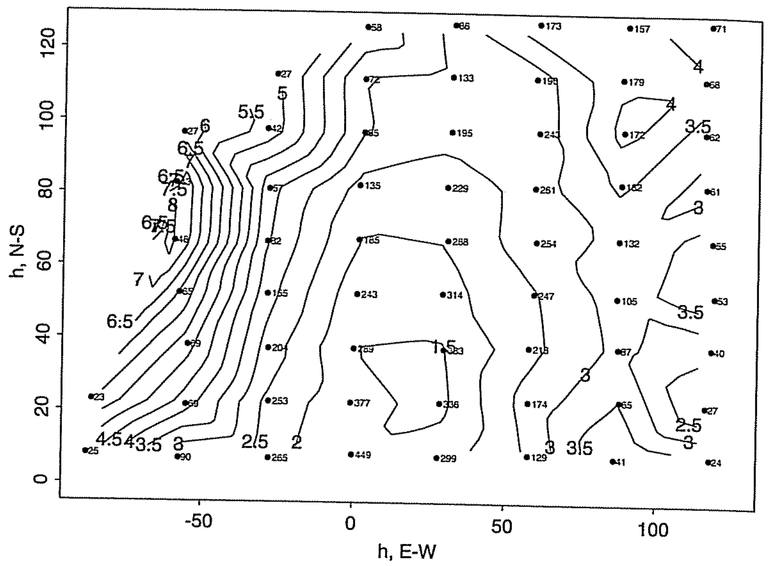


Figure 2.10 *ESC plot for the 1993 scallop data.*

of the angle and distance. For a chosen direction η and tolerance ϵ, the area for a class fans out as distance increases (see Figure 7.1 of Isaaks and Srivastava, 1989, p. 142). Attractively, a directional semivariogram based on the rectangular bins associated with the empirical semivariogram in $\Re^2$ has bin area remaining constant as distance increases. In Figure 2.11, we present the four customary directional (polar representation) semivariograms for the 1993 scallop data. Clearly, the ESC plot is more informative, particularly in suggesting evidence of geometric anisotropy.

2.4 Classical spatial prediction

In this section we describe the classical (i.e., minimum mean-squared error) approach to spatial prediction in the point-referenced data setting. The approach is commonly referred to as *kriging*, so named by Matheron (1963) in honor of D.G. Krige, a South African mining engineer whose seminal work on empirical methods for geostatistical data (Krige, 1951) inspired the general approach (and indeed, inspired the convention of using the terms "point-level spatial" and "geostatistical" interchangeably!). The problem is one of optimal spatial prediction: given observations of a random field $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$, how do we predict the variable Y at a site $\mathbf{s}_0$ where it has not been observed? In other words, what is the best predictor of the value of $Y(\mathbf{s}_0)$ based upon the data $\mathbf{y}$?

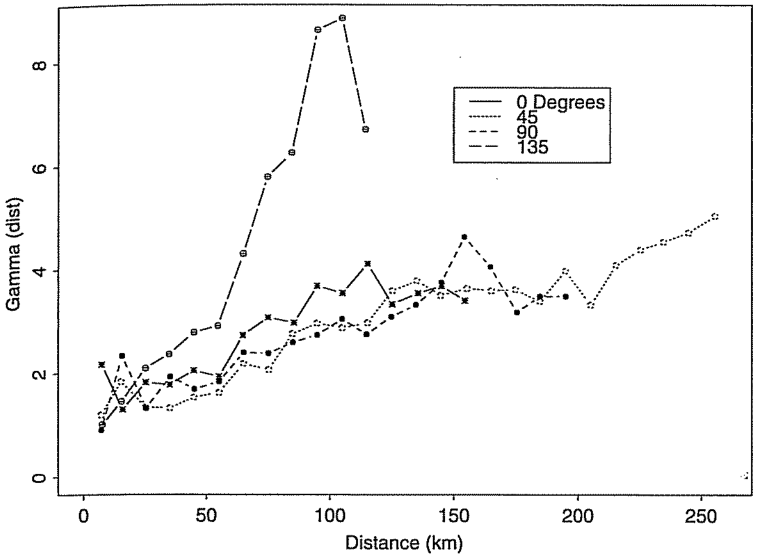


Figure 2.11 *Directional semivariograms for the 1993 scallop data.*

A linear predictor for $Y(\mathbf{s}_0)$ based on $\mathbf{y}$ would take the form $\sum \ell_i Y(\mathbf{s}_i) + \delta_0$. Using squared error loss, the best linear prediction would minimize $E[Y(\mathbf{s}_0) - (\sum \ell_i Y(\mathbf{s}_i) + \delta_0)]^2$ over δ_0 and the ℓ_i. For a constant mean process we would take $\sum \ell_i = 1$, in which case we would minimize $E[Y(\mathbf{s}_0) - \sum \ell_i Y(\mathbf{s}_i)]^2 + \delta_0^2$, and clearly δ_0 would be set to 0. Now letting $a_0 = 1$ and $a_i = -\ell_i$ we see that the criterion becomes $E[\sum_{i=0}^n a_i Y(\mathbf{s}_i)]^2$ with $\sum a_i = 0$. But from (2.4) this expectation becomes $-\sum_i \sum_j a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j)$, revealing how, historically, the variogram arose in kriging within the geostatistical framework. Indeed, the optimal ℓ's can be obtained by solving this constrained optimization (e.g., using Lagrange multipliers), and will be functions of $\gamma(\mathbf{h})$ (see, e.g., Cressie, 1983, Sec. 3.2). With an estimate of γ, one immediately obtains the so-called *ordinary kriging* estimate. Other than the intrinsic stationarity model (Subsection 2.1.2), no further distributional assumptions are required for the $Y(\mathbf{s})$'s.

Let us take a more formal look at kriging in the context of Gaussian processes. Consider first the case where we have no covariates, but only the responses $Y(\mathbf{s}_i)$. This is developed by means of the following model for the observed data:

$$\mathbf{Y} = \mu \mathbf{1} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma) \ .$$

For a spatial covariance structure having no nugget effect, we specify Σ as

$$\Sigma = \sigma^2 H(\phi) \text{ where } (H(\phi))_{ij} = \rho(\phi; d_{ij}) \ ,$$

where $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$, the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$ and ρ is a valid correlation function on $\Re^r$ such as those in Table 2.1. For a model having a nugget effect, we instead set

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I \,,$$

where τ^2 is the nugget effect variance.

When covariate values $\mathbf{x} = (x(\mathbf{s}_1), \ldots, x(\mathbf{s}_n))'$ and $x(\mathbf{s}_0)$ are available for incorporation into the analysis, the procedure is often referred to as *universal kriging*, though we caution that some authors (e.g., Kaluzny et al., 1998) use the term "universal" in reference to the case where only latitude and longitude are available as covariates. The model now takes the more general form

$$\mathbf{Y} = X\beta + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma) \,,$$

with Σ being specified as above, either with or without the nugget effect. Note that ordinary kriging may be looked upon as a particular case of universal kriging with X being the $n \times 1$ matrix (i.e., column vector) $\mathbf{1}$, and β the scalar μ.

We now pose our prediction problem as follows: we seek the function $f(\mathbf{y})$ that minimizes the mean-squared prediction error,

$$E\left[\left(Y(\mathbf{s}_0) - f(\mathbf{y})\right)^2 \middle| \mathbf{y}\right] . \tag{2.17}$$

By adding and subtracting the conditional mean $E[Y(\mathbf{s}_0)|\mathbf{y}]$ inside the square, grouping terms, and squaring we obtain

$$\begin{aligned} & E\left[\left(Y(\mathbf{s}_0) - f(\mathbf{y})\right)^2 \middle| \mathbf{y}\right] \\ & \quad = E\left\{\left(Y(\mathbf{s}_0) - E[Y(\mathbf{s}_0)|\mathbf{y}]\right)^2 \middle| \mathbf{y}\right\} + \left\{E[Y(\mathbf{s}_0)|\mathbf{y}] - f(\mathbf{y})\right\}^2 , \end{aligned}$$

since (as often happens in statistical derivations like this) the expectation of the cross-product term equals zero. But since the second term on the right-hand side is nonnegative, we have

$$E\left[\left(Y(\mathbf{s}_0) - f(\mathbf{y})\right)^2 \middle| \mathbf{y}\right] \geq E\left\{\left(Y(\mathbf{s}_0) - E[Y(\mathbf{s}_0)|\mathbf{y}]\right)^2 \middle| \mathbf{y}\right\}$$

for any function $f(\mathbf{y})$. Equality holds if and only if $f(\mathbf{y}) = E[Y(\mathbf{s}_0)|\mathbf{y}]$, so it must be that the predictor $f(\mathbf{y})$ that minimizes the error is the conditional expectation of $Y(\mathbf{s}_0)$ given the data. This result is quite intuitive from a Bayesian point of view, since this $f(\mathbf{y})$ is just the *posterior mean* of $Y(\mathbf{s}_0)$, and it is well known that the posterior mean is the Bayes rule (i.e., the minimizer of posterior risk) under squared error loss functions of the sort adopted in (2.17) above as our scoring rule.

Having identified the form of the best predictor we now turn to its estimation. Consider first the wildly unrealistic situation in which all the population parameters (β, σ^2, ϕ, and τ^2) are known. From standard multivariate normal theory we have the following general result: If

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}\right),$$

where $\Omega_{21} = \Omega_{12}^T$, then the conditional distribution $p(\mathbf{Y}_1|\mathbf{Y}_2)$ is normal with mean and variance:

$$\begin{aligned} E[\mathbf{Y}_1|\mathbf{Y}_2] &= \boldsymbol{\mu}_1 + \Omega_{12}\Omega_{22}^{-1}(\mathbf{Y}_2 - \boldsymbol{\mu}_2) ; \\ Var[\mathbf{Y}_1|\mathbf{Y}_2] &= \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21} . \end{aligned}$$

In our framework, we have $\mathbf{Y}_1 = Y(\mathbf{s}_0)$ and $\mathbf{Y}_2 = \mathbf{y}$. It then follows that

$$\Omega_{11} = \sigma^2 + \tau^2, \quad \Omega_{12} = \boldsymbol{\gamma}^T, \quad \text{and } \Omega_{22} = \Sigma = \sigma^2 H(\phi) + \tau^2 I \,,$$

where $\boldsymbol{\gamma}^T = \left(\sigma^2\rho(\phi; d_{01}), \ldots, \sigma^2\rho(\phi; d_{0n})\right)$. Substituting these values into the mean and variance formulae above, we obtain

$$\begin{aligned} E[Y(\mathbf{s}_0)|\mathbf{y}] &= \mathbf{x}_0^T\beta + \boldsymbol{\gamma}^T\Sigma^{-1}(\mathbf{y} - X\beta) \,, & (2.18) \\ \text{and } Var[Y(\mathbf{s}_0)|\mathbf{y}] &= \sigma^2 + \tau^2 - \boldsymbol{\gamma}^T\Sigma^{-1}\boldsymbol{\gamma} \,. & (2.19) \end{aligned}$$

We remark that this solution assumes we have actually observed the covariate value $\mathbf{x}_0 = \mathbf{x}(\mathbf{s}_0)$ at the "new" site $\mathbf{s}_0$; we defer the issue of missing $\mathbf{x}_0$ for the time being.

Note that one could consider prediction *not* at a new location, but at one of the already observed locations. In this case one can ask whether or not the predictor in (2.18) will equal the observed value at that location. We leave it as an exercise to verify that if $\tau^2 = 0$ (i.e., the no-nugget case, or so-called noiseless prediction) then the answer is yes, while if $\tau^2 > 0$ then the answer is no.

Next, consider how these answers are modified in the more realistic scenario where the model parameters are unknown and so must be estimated from the data. Here we would modify $f(\mathbf{y})$ to

$$\widehat{f(\mathbf{y})} = \mathbf{x}_0^T\widehat{\beta} + \widehat{\boldsymbol{\gamma}}^T\widehat{\Sigma}^{-1}\left(\mathbf{y} - X\widehat{\beta}\right) ,$$

where $\widehat{\boldsymbol{\gamma}} = \left(\hat{\sigma}^2\rho(\hat{\phi}; d_{01}), \ldots, \hat{\sigma}^2\rho(\hat{\phi}; d_{0n})\right)^T$, $\widehat{\beta} = \left(X^T\widehat{\Sigma}^{-1}X\right)^{-1}X^T\widehat{\Sigma}^{-1}\mathbf{y}$, the usual weighted least squares estimator of β, and $\widehat{\Sigma} = \hat{\sigma}^2 H(\hat{\phi})$. Thus $\widehat{f(\mathbf{y})}$ can be written as $\boldsymbol{\lambda}^T\mathbf{y}$, where

$$\boldsymbol{\lambda} = \widehat{\Sigma}^{-1}\widehat{\boldsymbol{\gamma}} + \widehat{\Sigma}^{-1}X\left(X^T\widehat{\Sigma}^{-1}X\right)^{-1}\left(\mathbf{x}_0 - X^T\widehat{\Sigma}^{-1}\widehat{\boldsymbol{\gamma}}\right) . \tag{2.20}$$

If $\mathbf{x}_0$ is unobserved, we can estimate it and $Y(\mathbf{s}_0)$ jointly by iterating between this formula and a corresponding one for $\hat{\mathbf{x}}_0$, namely

$$\hat{\mathbf{x}}_0 = X^T\boldsymbol{\lambda} \,,$$

which arises simply by multiplying both sides of (2.20) by X^T and simplifying. This is essentially an EM (expectation-maximization) algorithm (Dempster, Laird, and Rubin, 1977), with the calculation of $\hat{\mathbf{x}}_0$ being the E step and (2.20) being the M step.

In the classical framework a lot of energy is devoted to the determination of the optimal estimates to plug into the above equations. Typically, restricted maximum likelihood (REML) estimates are selected and shown to have certain optimal properties. However, as we shall see in Chapter 5, how to perform the estimation is not an issue in the Bayesian setting. There, we instead impose prior distributions on the parameters and produce the full posterior predictive distribution $p\left(Y(\mathbf{s}_0)|\mathbf{y}\right)$. Any desired point or interval estimate (the latter to express our uncertainty in such prediction) may then be computed with respect to this distribution.

2.5 Computer tutorials

2.5.1 EDA and variogram fitting in S+SpatialStats

In this section we outline the use of the S+SpatialStats package in performing exploratory analysis on spatially referenced data. Throughout we use a "computer tutorial" style, as follows.

First, we need to load the spatial module into the S-plus environment:

```
>module(spatial)
```

The scallops data, giving locations and scallop catches in the Atlantic waters off the coasts of New Jersey and Long Island, New York, is preloaded as a data frame in S-plus, and can therefore be accessed directly. For example, a descriptive summary of the data can be obtained by typing

```
>summary(scallops)
```

In order to present graphs and maps in S-plus we will need to open a graphing device. The best such device is called trellis.device(). Here we draw a histogram of the variable tcatch in the dataframe scallops. Note the generic notation a$b for accessing a member b of a dataframe a. Thus the member tcatch of dataframe scallops is accessed as scallops$tcatch. We also print the histogram to a .ps file:

```
>trellis.device()
>hist(scallops$tcatch)
>printgraph(file=‘‘histogram.tcatch.ps’’)
```

Noticing the data to be highly skewed, we feel the need to create a new variable log(tcatch). But since tcatch contains a number of 0's, we instead compute *log*(*tcatch* + 1). For that it is best to create our own dataframe since it is not a good idea to "spoil" S-plus' own dataframe. As such we assign.scallops to myscallops. Then we append the variable lgcatch

(which is actually *log*(*tcatch* + 1)) to myscallops. We then draw the histogram of the variable lgcatch:

```
>myscallops <- scallops
>myscallops[,‘‘lgcatch’’] <- log(scallops$tcatch+1)
>summary(myscallops$lgcatch)
>hist(myscallops$lgcatch)
```

This histogram exhibits much more symmetry than the earlier one, suggesting a normality assumption we might make later when kriging will be easier to accept.

We next plot the locations as an ordinary line plot:

```
>plot(myscallops$long, myscallops$lat)
```

For spatial purposes, we actually need a *geographic information system (GIS)* interface. This is offered by the S-plus library maps. We next invoke this library and extract a map of the U.S. from it.

```
>library(maps)
>map(‘‘usa’’)
```

For our data, however, we do not need the map of the entire U.S. Looking at the earlier summary of scallops, we note the range of the latitude and longitude variables and decide upon the following limits. Note that xlim sets the x-axis limits and ylim sets the y-axis limits; the c() function creates vectors.

```
>map(‘‘usa’’, xlim=c(-74, -71), ylim=c(38.2, 41.5))
```

The observed sites may be embedded on the map, reducing their size somewhat using the cex ("character expansion") option:

```
>points(myscallops$long, myscallops$lat, cex=0.75)
```

It is often helpful to add contour lines to the plot. In order to add such lines it is necessary to carry out an interpolation. This essentially fills in the gaps in the data over a regular grid (where there are no actual observerd data) using a bivariate linear interpolation. This is done in S-plus using the interp function. The contour lines may then be added to the plot using the contour command:

```
>int.scp <- interp(myscallops$long,
   myscallops$lat, myscallops$lgcatch)
>contour(int.scp, add=T)
```

Figure 2.12 shows the result of the last four commands, i.e., the map of the scallop locations and log catch contours arising from the linear interpolation.

Two other useful ways of looking at the data may be through image and perspective (three-dimensional surface) plots. Remember that they will use the interpolated object so a preexisting interpolation is also compulsory here.

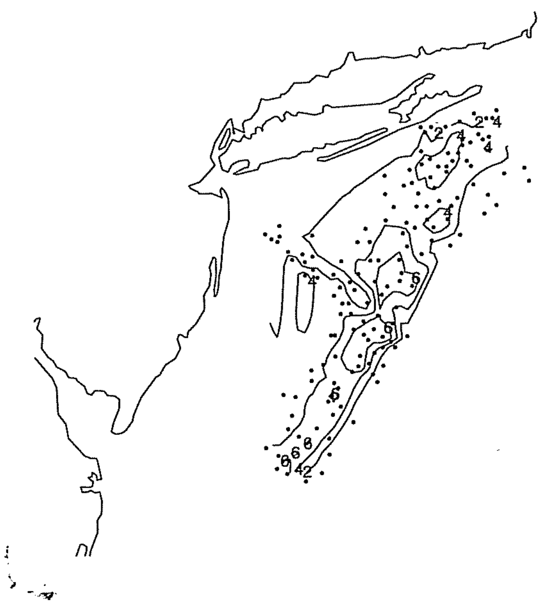


Figure 2.12 *Map of observed scallop sites and contours of (linearly interpolated) raw log catch data, scallop data.*

```
>image(int.scp)
>persp(int.scp)
```

The empirical variogram can be estimated in both the standard and "robust" (Cressie and Hawkins) way with built-in functions. We first demonstrate the standard approach. After a variogram object is created, typing that object yields the actual values of the variogram function with the distances at which they are computed. A summary of the object may be invoked to see information for each lag, the total number of lags, and the maximum intersite distance.

```
>scallops.var <- variogram(lgcatch~loc(long,lat),
   data=myscallops)
>scallops.var
>summary(scallops.var)
```

In `scallops.var`, `distance` corresponds to the spatial lag (h in our usual notation), `gamma` is the variogram $\gamma(h)$, and `np` is the number of points in each bin. In the output of the `summary` command, `maxdist` is the largest distance on the map, `nlag` is the number of lags (variogram bins), and `lag` is `maxdist/nlag`, which is the width of each variogram bin.

By contrast, the robust method is obtained simply by specifying "robust" in the method option:

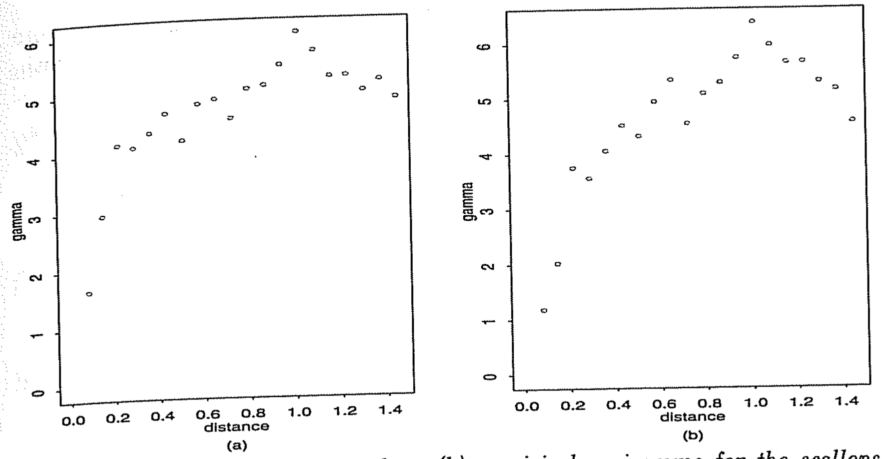


Figure 2.13 *Ordinary (a) and robust (b) empirical variograms for the scallops data.*

```
>scallops.var.robust <- variogram(lgcatch~loc(long,lat),
   data=myscallops, method = ''robust'')
```

Plotting is usually done by just calling the `plot` function with the variogram object as its argument. It may be useful to compare the plots one below the other. Setting up these plots is done as follows:

```
>par(mfrow=c(1,2))
>plot(scallops.var)
>plot(scallops.var.robust)
>printgraph(file=''scallops.empvario.ps'')
```

The output from this picture is shown in Figure 2.13.

The *covariogram* (a plot of an isotropic empirical covariance function (2.15) versus distance) and *correlogram* (a plot of (2.15) divided by $\hat{C}(\mathbf{0})$ versus distance) may be created using the `covariogram` and `correlogram` functions. (When we are through here, we set the graphics device back to having one plot per page using the `par` command.)

```
>scallops.cov <- covariogram(lgcatch~loc(long,lat),
   data=myscallops)
>plot(scallops.cov)
>scallops.corr <- correlogram(lgcatch~loc(long,lat),
   data=myscallops)
>plot(scallops.corr)
>printgraph(file=''scallops.covariograms.ps'')
> par(mfrow=c(1,1))
```

Theoretical variograms may also be computed and compared to the observed data as follows. Invoke the `model.variogram` function and choose an initial theoretical model; say, range=0.8, sill=1.25, and nugget=0.50. Note that the fun option specifies the variogram type we want to work with. Below we choose the spherical (spher.vgram); other options include exponential (exp.vgram), Gaussian (gauss.vgram), linear (linear.vgram), and power (power.vgram).

```
>model.variogram(scallops.var.robust, fun=spher.vgram,
   range=0.80, sill=1.25, nugget = 0.50)
```

We remark that this particular model provides relatively poor fit to the data; the objective function takes a relatively high value (roughly 213). (You are asked to find a better-fitting model in Exercise 7.)

Formal estimation procedures for variograms may also be carried out by invoking the nls function on the spher.fun function that we can create:

```
>spher.fun <- function(gamma,distance,range,sill,nugget){
    gamma - spher.vgram(distance, range=range,
   sill=sill, nugget=nugget)}
>scallops.nl1 <- nls(~spher.fun(gamma, distance, range,
   sill, nugget), data = scallops.var.robust,
   start=list(range=0.8, sill=1.05, nugget=0.7))
>coef(scallops.nl1)
```

Thus we are using nls to minimize the squared distance between the theoretical and empirical variograms. Note there is nothing to the left of the "~" character at the beginning of the nls statement.

Many times our interest lies in spatial *residuals*, or what remains after detrending the response from the effects of latitude and longitude. An easy way to do that is by using the gam function in S-plus. Here we plot the residuals of the scallops lgcatch variable after the effects of latitude and longitude have been accounted for:

```
>gam.scp <- gam(lgcatch~lo(long)+lo(lat), data= myscallops)
>par(mfrow=c(2,1))
>plot(gam.scp, residuals=T, rug=F)
```

Finally, at the end of the session we unload the spatial module, after which we can either do other work, or quit.

```
>module(spatial, unload=T)
>q()
```

2.5.2 *Kriging in* S+SpatialStats

We now present a tutorial in using S+SpatialStats to do basic kriging. At the command prompt type S-plus to start the software, and load the spatial module into the environment:

```
>module(spatial)
```

Recall that the scallops data is preloaded as a data frame in S-plus, and a descriptive summary of this data set can be obtained by typing

```
>summary(scallops)
```

while the first row of the data may be seen by typing

```
>scallops[1,]
```

Recall from our Section 2.5.1 tutorial that the data on tcatch was highly skewed, so we needed to create another dataframe called myscallops, which includes the log transform of tcatch (or actually log(*tcatch*+1)). We called this new variable lgcatch. We then computed both the regular empirical variogram and the "robust" (Cressie and Hawkins) version, and compared both to potential theoretical models using the variogram command.

```
>scallops.var.robust <- variogram(lgcatch~loc(long,lat),
   data=myscallops, method = ``robust'')
```

Plotting is usually done using the plot function on the variogram object:

```
>trellis.device()
>plot(scallops.var.robust)
```

Next we recall S-plus' ability to compute theoretical variograms. We invoke the model.variogram function, choosing a theoretical starting model (here, range=0.8, sill=4.05, and nugget=0.80), and using fun to specify the variogram type.

```
>model.variogram(scallops.var.robust, fun=spher.vgram,
   range=0.80, sill=4.05, nugget = 0.80)
>printgraph(file=``scallops.variograms.ps'')
```

The output from this command (robust empirical semivariogram with this theoretical variogram overlaid) is shown in Figure 2.14. Note again the model.variogram command allows the user to alter the theoretical model and continually recheck the value of the objective function (where smaller values indicate better fit of the theoretical to the empirical).

Formal estimation procedures for variograms may also be carried out by invoking the nls function on the spher.fun function that we create:

```
>spher.fun <- function(gamma,distance,range,sill,nugget){
    gamma - spher.vgram(distance, range=range,
   sill=sill, nugget=nugget)}
>scallops.nl1 <- nls(~spher.fun(gamma, distance, range,
   sill, nugget), data = scallops.var.robust;
   start=list(range=0.8, sill=4.05, nugget=0.8))
>summary(scallops.nl1)
```

We now call the kriging function krige on the variogram object to produce estimates of the parameters for ordinary kriging:

```
>scallops.krige <- krige(lgcatch~loc(long,lat),
   data=myscallops, covfun=spher.cov, range=0.71,
```

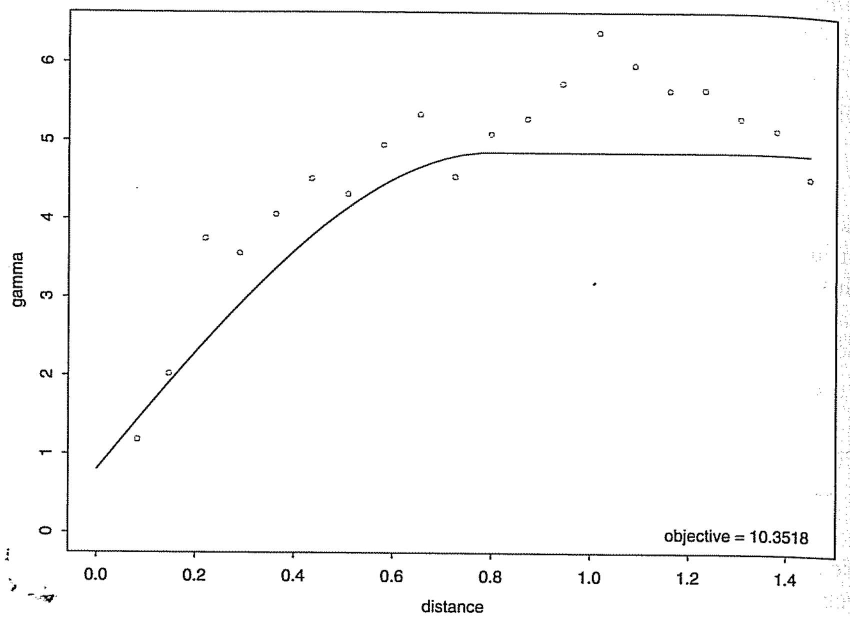


Figure 2.14 *Robust empirical and theoretical (spherical) variograms for the scallops data.*

```
    nugget=0.84, sill=4.53)
```

Note that the covfun option here specifies an intrinsic spherical covariance function. Now suppose we want to predict the response at a small collection of new locations. We need to create a new text file containing the latitudes and longitudes for these new locations. We *must* label the coordinates as lat and long, exactly matching the names in our original dataframe myscallops. Download the file

www.biostat.umn.edu/~brad/data/newdata.txt

from the web, and save the file as newdata.txt. The file contains two new locations:

```
  long   lat
-71.00  40.0
-72.75  39.5
```

Here the first site is far from the bulk of the observed data (so the predicted values should have high standard errors), while the second site is near the bulk of the observed data (so the predicted values should have low standard errors).

Next we create a data frame called newdata that reads in the new set of sites using the read.table function in S-plus, remembering to include the header=T option. Having done that we may call the predict function, specifying our newdata data frame using the newdata option:

```
>newdata <- read.table(''newdata.txt'', header=T)
>scallops.predicttwo <- predict(scallops.krige,
   newdata=newdata)
```

scallops.predicttwo then contains the predictions and associated standard errors for the two new sites, the latter of which are ordered as we anticipated.

Next, we consider the case where we wish to predict not at a few specific locations, but over a fine grid of sites, thus enabling a prediction *surface*. In such a situation one can use the expand.grid function, but the interp function seems to offer an easier and less error-prone approach. To do this, we first call the predict function *without* the newdata option. After checking the first row of the scallops.predict object, we collect the coordinates and the predicted values into three vectors x, y, and z. We then invoke the interp and persp functions:

```
>scallops.predict <- predict(scallops.krige)
>scallops.predict[1,]
>x <- scallops.predict[,1]
>y <- scallops.predict[,2]
>z <- scallops.predict[,3]
>scallops.predict.interp <- interp(x,y,z)
>persp(scallops.predict.interp)
```

It may be useful to recall the location of the sites and the surface plot of the raw data for comparison. We create these plots on a separate graphics device:

```
>trellis.device()
>plot(myscallops$long, myscallops$lat)
>int.scp <- interp(myscallops$long, myscallops$lat,
   myscallops$lgcatch)
>persp(int.scp)
```

Figure 2.15 shows these two perspective plots side by side for comparison. The predicted surface on the left is smoother, as expected.

It is also useful to have a surface plot of the standard errors, since we expect to see higher standard errors where there is less data. This is well illustrated by the following commands:

```
>z.se <- scallops.predict[,4]
>scallops.predict.interp.se <- interp(x,y,z.se)
>persp(scallops.predict.interp.se)
```

Other plots, such as image plots for the prediction surface with added contour lines, may be useful:

```
>image(scallops.predict.interp)
```

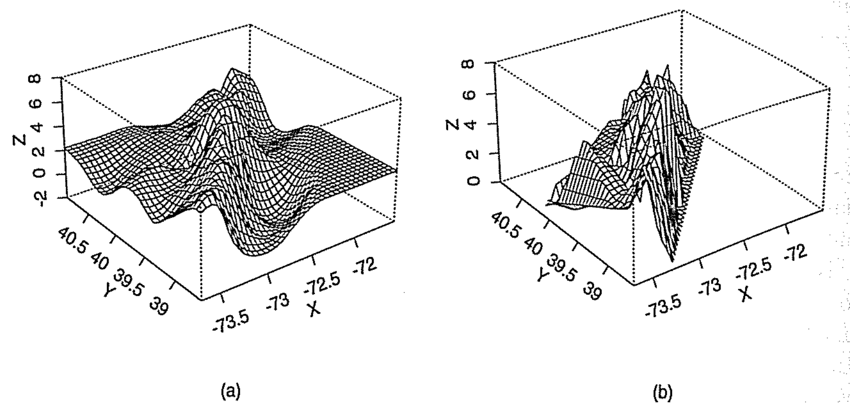


Figure 2.15 *Perspective plots of the kriged prediction surface (a) and interpolated raw data (b), log scallop catch data.*

```
>par(new=T, xaxs=‘‘d’’, yaxs=‘‘d’’)
>contour(scallops.predict.interp)
```

Turning to universal kriging, here we illustrate with the scallops data using latitude and longitude as the covariates (i.e., trend surface modeling). Our covariance matrix X is therefore $n \times 3$ ($n = 148$ here) with columns corresponding to the intercept, latitude and longitude.

```
>scallops.krige.universal <- krige(lgcatch~loc(long,lat)
   +long+lat, data=myscallops, covfun=spher.cov,
   range=0.71, nugget=0.84, sill=4.53)
```

Note that the scallops.krige.universal function gives point estimates, but not associated standard errors. You are asked to remedy this situation in Exercise 11.

Plots like those already seen for ordinary kriging may be done as well. It is also useful to produce a spatial surface of the standard errors of the fit.

```
>scallops.predict.universal
   <- predict(scallops.krige.universal)
>scallops.predict.universal[1,]
>x <- scallops.predict.universal[,1]
>y <- scallops.predict.universal[,2]
>z <- scallops.predict.universal[,3]
>scallops.predict.interp <- interp(x,y,z)
>persp(scallops.predict.interp)
>q()
```

2.5.3 EDA, variograms, and kriging in geoR

R is an increasing popular freeware alternative to S-plus, available from the web at www.r-project.org. In this subsection we describe methods for kriging and related geostatistical operations available in geoR, a geostatistical data analysis package using R, which is also freely available on the web at www.est.ufpr.br/geoR/.

Since the syntax of S-plus and R is virtually identical, we do not spend time here repeating the material of the past subsection. Rather, we only highlight a few differences in exploratory data analysis steps, before moving on to model fitting and kriging.

Consider again the scallop data. We recall that it is often helpful to create image plots and place contour lines on the plot. These provide a visual idea of the realized spatial surface. In order to do these, it is necessary to first carry out an interpolation. This essentially fills up the gaps (i.e., where there are no points) using a bivariate linear interpolation. This is done using the interp.new function in R, located in the library akima. Then the contour lines may be added to the plot using the contour command. The results are shown in Figure 2.16.

```
>library(akima)
>int.scp_interp.new(myscallops$long, myscallops$lat,
   myscallops$lgcatch, extrap=T)
>image(int.scp, xlim=range(myscallops$long),
   ylim=range(myscallops$lat))
>contour(int.scp, add=T)
```

Another useful way of looking at the data is through surface plots (or perspective plots). This is done by invoking the persp function:

```
>persp(int.scp, xlim=range(myscallops$long),
   ylim=range(myscallops$lat))
```

The empirical variogram can be estimated in the classical way and in the robust way with in-built R functions. There are several packages in R that perform the above computations. We illustrate the geoR package, mainly because of its additional ability to fit Bayesian geostatistical models as well. Nevertheless, the reader might want to check out the CRAN website (http://cran.us.r-project.org/) for the latest updates and several other spatial packages. In particular, we mention fields, gstat, sgeostat, spatstat, and spatdep for exploratory work and some model fitting of spatial data, and GRASS and RArcInfo for interfaces to GIS software.

Returning to the problem of empirical variogram fitting, we first invoke the geoR package. We will use the function variog in this package, which takes in a geodata object as input. To do this, we first create an object, obj, with only the coordinates and the response. We then create the geodata

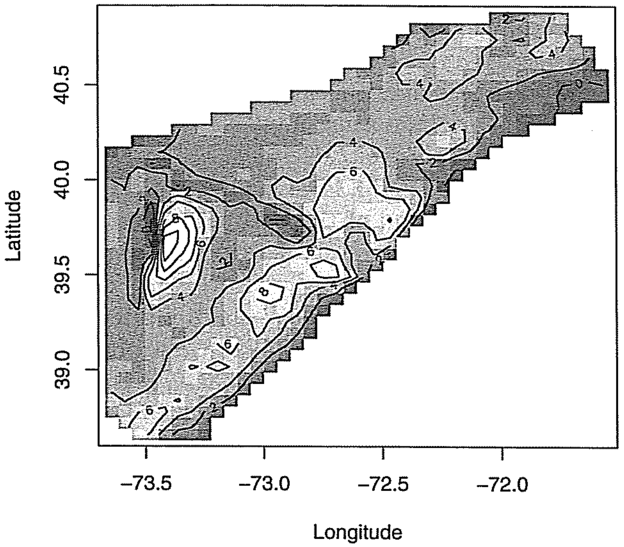


Figure 2.16 *An image plot of the scallops data, with contour lines super-imposed.*

object using the as.geodata function, specifying the columns holding the coordinates, and the one holding the response.

```
>obj_cbind(myscallops$long,myscallops$lat,
   myscallops$lgcatch)
>scallops.geo_as.geodata(myscallops,coords.col=1:2,
   data.col=3)
```

Now, a variogram object is created.

```
>scallops.var_variogram(scallops.geo,
   estimator.type=''classical'')
>scallops.var
```

The robust estimator (see Cressie, 1993, p.75) can be obtained by typing

```
>scallops.var.robust_variogram(scallops.geo,
   estimator.type=''modulus'')
```

A plot of the two semivariograms (by both methods, one below the other, as in Figure 2.17) can be obtained as follows:

```
>par(mfrow=c(2,1))
>plot(scallops.var)
>plot(scallops.var.robust)
```

Covariograms and correlograms are invoked using the covariogram and correlogram functions. The remaining syntax is the same as in S-plus.

The function variofit estimates the sill, the range, and the nugget

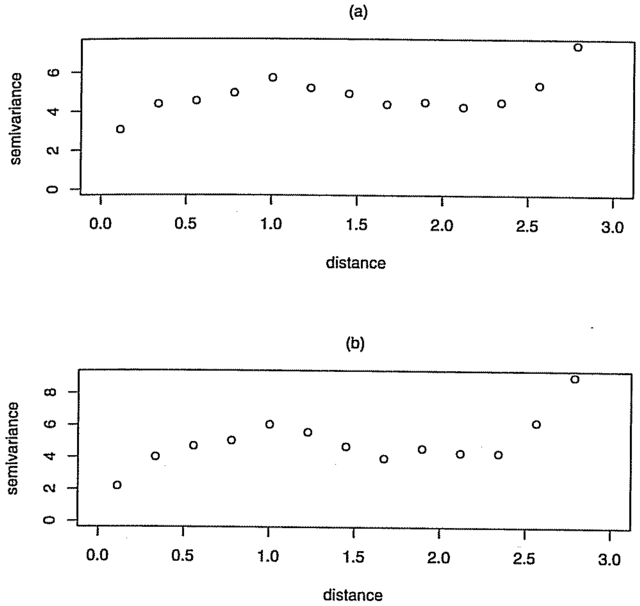


Figure 2.17 *Plots of the empirical semivariograms for the scallops data: (a) classical; (b) robust.*

parameters under a specified covariance model. A variogram object (typically an output from the variog function) is taken as input, together with initial values for the range and sill (in ini.cov.pars), and the covariance model is specified through cov.model. The covariance modeling options include exponential, gaussian, spherical, circular, cubic, wave, power, powered.exponential, cauchy, gneiting, gneiting.matern, and pure.nugget (no spatial covariance). Also, the initial values provided in ini.cov.pars do not include those for the nugget. It is concatenated with the value of the nugget option only if fix.nugget=FALSE. If the latter is TRUE, then the value in the nugget option is taken as the fixed true value.

Thus, with the exponential covariance function for the scallops data, we can estimate the parameters (including the nugget effect) using

```
>scallops.var.fit_variofit(scallops.var.robust,
   ini.cov.pars = c(1.0,50.0), cov.model=''exponential'',
   fix.nugget=FALSE, nugget=1.0)
```

The output is given below. Notice that this is the weighted least squares approach for fitting the variogram:

```
         variofit: model parameters estimated by WLS
                   (weighted least squares):
      covariance model is: matern with fixed kappa = 0.5
                         (exponential)
```

```
                parameter estimates:
                 tausq sigmasq phi
                0.0000 5.1289 0.2160
```

Likelihood model fitting

In the previous section we saw parameter estimation through weighted least squares of variograms. Now we introduce likelihood-based and Bayesian estimation functions in geoR.

Both maximum likelihood and REML methods are available through the geoR function likfit. To estimate the parameters for the scallops data, we invoke

```
>scallops.lik.fit_likfit(scallops.geo,
   ini.cov.pars=c(1.0,2.0),cov.model = ‘‘exponential’’,
   trend = ‘‘cte’’, fix.nugget = FALSE, nugget = 1.0,
   nospatial = TRUE, method.lik = ‘‘ML’’)
```

The option trend = ‘‘cte’’ means a spatial regression model with constant mean. This yields the following output:

```
> scallops.lik.fit

            likfit: estimated model parameters:
                   beta tausq sigmasq phi
                2.3748 0.0947 5.7675 0.2338
```

Changing method.lik = ‘‘REML’’ yields the restricted maximum likelihood estimation. Note that the variance of the estimate of beta is available by invoking scallops.lik.fit$beta.var, so calculating the confidence interval for the trend is easy. However, the variances of the estimates of the covariance parameters is not easily available within geoR.

Kriging in geoR

There are two in-built functions in geoR for kriging: one is for classical or conventional kriging, and is called krige.conv, while the other performs Bayesian kriging and is named krige.bayes. We now briefly look into these two types of functions. The krige.bayes function is not as versatile as WinBUGS in that it is more limited in the types of models it can handle, and also the updating is not through MCMC methods. Nevertheless, it is a handy tool and already improved upon the aforementioned likelihood methods by providing posterior samples of *all* the model parameters, which lead to estimation of their variability.

The krige.bayes function can be used to estimate parameters for spatial regression models. To fit a constant mean spatial regression model for the scallops data, without doing predictions, we invoke krige.bayes specifying a constant trend, an exponential covariance model, a flat prior for the constant trend level, the reciprocal prior for sigmasq (Jeffrey's), and a discrete uniform prior for tausq.

```
>scallops.bayes1 <- krige.bayes(scallops.geo,
   locations = ‘‘no’’, borders = NULL, model =
   model.control(trend.d = ‘‘cte’’,
   cov.model = ‘‘exponential’’),
   prior = prior.control(beta.prior = ‘‘flat’’,
   sigmasq.prior = ‘‘reciprocal’’,
   tausq.rel.prior = ‘‘uniform’’,
   tausq.rel.discrete=seq(from=0.0,to=1.0,by=0.01)))
```

We next form the quantiles in the following way:

```
> out_scallops.krige.bayes$posterior
> out_out$sample
> beta.qnt_quantile(out$beta, c(0.50,0.025,0.975))
> phi.qnt_quantile(out$phi, c(0.50,0.025,0.975))
> sigmasq.qnt_quantile(out$sigmasq, c(0.50,0.025,0.975))
> tausq.rel.qnt_quantile(out$tausq.rel,
  c(0.50,0.025,0.975))
> beta.qnt

                   50% 2.5% 97.5%
             1.931822 -6.426464 7.786515

> phi.qnt

                   50% 2.5% 97.5%
             0.5800106 0.2320042 4.9909913

sigmasq.qnt

                   50% 2.5% 97.5%
             11.225002 4.147358 98.484722

> tausq.rel.qnt

                   50% 2.5% 97.5%
                   0.03 0.00 0.19
```

Note that tausq.rel refers to the ratio of the nugget variance to the spatial variance, and is seen to be negligible here, too. This is consistent with all the earlier analysis, showing that a purely spatial model (no nugget) would perhaps be more suitable for the scallops data.

2.6 Exercises

1. For semivariogram models #2, 4, 5, 6, 7, and 8 in Subsection 2.1.3,

   (a) identify the nugget, sill, and range (or effective range) for each;

(b) find the covariance function $C(t)$ corresponding to each $\gamma(t)$, provided it exists.

2. Prove that for Gaussian processes, strong stationarity is equivalent to weak stationarity.

3. Consider the *triangular* (or "tent") covariance function,

$$C(\|h\|) = \begin{cases} \sigma^2(1-\|h\|/\delta) & \text{if } \|h\| \le \delta,\ \sigma^2 > 0,\ \delta > 0, \\ 0 & \text{if } \|h\| > \delta \end{cases} .$$

It is valid in one dimension. (The reader can verify that it is the characteristic function of the density function $f(x)$ proportional to $1 - \cos(\delta x)/\delta x^2$.) Now in two dimensions, consider a 6×8 grid with locations $\mathbf{s}_{jk} = (j\delta/\sqrt{2}, k\delta/\sqrt{2})$, $j = 1, \ldots, 6$, $k = 1, \ldots, 8$. Assign a_{jk} to $\mathbf{s}_{jk}$ such that $a_{jk} = 1$ if $j+k$ is even, $a_{jk} = -1$ if $j+k$ is odd. Show that $Var[\Sigma a_{jk} Y(\mathbf{s}_{jk})] < 0$, and hence that the triangular covariance function is *invalid* in two dimensions.

4. The *turning bands method* (Christakos, 1984; Stein, 1999a) is a technique for creating stationary covariance functions on $\Re^r$. Let $\mathbf{u}$ be a random unit-vector on $\Re^r$ (by random we mean that the coordinate vector that defines $\mathbf{u}$ is randomly chosen on the surface of the unit sphere in $\Re^r$). Let $c(\cdot)$ be a valid stationary covariance function on $\Re^1$, and let $W(t)$ be a process on $\Re^1$ having $c(\cdot)$ as its covariance function. Then for any location $\mathbf{s} \in \Re^r$, define

$$Y(\mathbf{s}) = W(\mathbf{s}^T \mathbf{u}) .$$

Note that we can think of the process either conditionally given $\mathbf{u}$, or marginally by integrating with respect to the uniform distribution for $\mathbf{u}$. Note also that $Y(\mathbf{s})$ has the possibly undesirable property that it is constant on planes (i.e., on $\mathbf{s}^T\mathbf{u} = k$).

(a) If W is a Gaussian process, show that, given $\mathbf{u}$, $Y(\mathbf{s})$ is also a Gaussian process and is stationary.

(b) Show that marginally $Y(\mathbf{s})$ is *not* a Gaussian process, but is isotropic. [*Hint:* Show that $Cov(Y(\mathbf{s}), Y(\mathbf{s}')) = E_{\mathbf{u}} c((\mathbf{s}-\mathbf{s}')^T\mathbf{u})$.]

5.(a) Based on (2.10), show that $c_{12}(\mathbf{h})$ is a valid correlation function; i.e., that G is a bounded, positive, symmetric about 0 measure on $\Re^2$.

(b) Show further that if c_1 and c_2 are isotropic, then c_{12} is.

6.(a) What is the issue with regard to specifying $\widehat{c}(0)$ in the covariance function estimate (2.15)?

(b) Show either algebraically or numerically that regardless of how $\widehat{c}(0)$ is obtained, $\widehat{\gamma}(t_k) \neq \widehat{c}(0) - \widehat{c}(t_k)$ for all t_k.

7. Carry out the steps outlined in Section 2.5.1 in S+SpatialStats. In addition:

(a) Provide a descriptive summary of the `scallops` data with the plots derived from the above session.

(b) Experiment with the `model.variogram` function to obtain rough estimates of the nugget, sill, and range; your final objective function should have a value less than 9.

(c) Repeat the theoretical variogram fitting with an exponential variogram, and report your results.

8. Consider the `coal.ash` data frame built into S+SpatialStats. This data comes from the Pittsburgh coal seam on the Robena Mine Property in Greene County, PA (Cressie, 1993, p. 32). This data frame contains 208 coal ash core samples (the variable `coal` in the data frame) collected on a grid given by x and y planar coordinates (*not* latitude and longitude). Carry out the following tasks in S-plus:

(a) Plot the sampled sites embedded on a map of the region. Add contour lines to the plot.

(b) Provide a descriptive summary (histograms, stems, quantiles, means, range, etc.) of the variable `coal` in the data frame.

(c) Plot variograms and correlograms of the response and comment on the need for spatial analysis here.

(d) If you think that there is need for spatial analysis, use the interactive `model.variogram` method in S-plus to arrive at your best estimates of the range, nugget, and sill. Report your values of the objective functions.

(e) Try to estimate the above parameters using the `nls` procedure in S-plus.

Hint: You may wish to look at Section 3.2 in Kaluzny et al. (1998) for some insight into the `coal.ash` data.

9. Confirm expressions (2.18) and (2.19), and subsequently verify the form for $\boldsymbol{\lambda}$ given in equation (2.20).

10. Show that when using (2.18) to predict the value of the surface at one of the existing data locations $\mathbf{s}_i$, the predictor will equal the observed value at that location if and only if $\tau^2 = 0$. (That is, the usual Gaussian process is a spatial interpolator only in the "noiseless prediction" scenario.)

11. It is an unfortunate feature of S+SpatialStats that there is no intrinsic routine to automatically obtain the standard errors of the estimated regression coefficients in the universal kriging model. Recall that

$$\begin{aligned} \mathbf{Y} &= X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma) , \\ \text{and } \Sigma &= \sigma^2 H(\phi) + \tau^2 I , \text{ where } (H(\phi))_{ij} = \rho(\phi; d_{ij}) . \end{aligned}$$

Thus the dispersion matrix of $\widehat{\boldsymbol{\beta}}$ is given as $Var(\widehat{\boldsymbol{\beta}}) = \left(X^T \Sigma^{-1} X\right)^{-1}$.

Thus $\widehat{Var}(\widehat{\beta}) = \left(X^T\widehat{\Sigma}^{-1}X\right)^{-1}$ where $\widehat{\Sigma} = \widehat{\sigma}^2 H(\widehat{\phi}) + \widehat{\tau}^2 I$ and $X = [\mathbf{1}, \texttt{long}, \texttt{lat}]$. Given the estimates of the sill, range, and nugget (from the `nls` function), it is possible to estimate the covariance matrix $\widehat{\Sigma}$, and thereby get $\widehat{Var}(\widehat{\beta})$. Develop an `S-plus` or `R` program to perform this exercise to obtain estimates of standard errors for $\widehat{\beta}$ for the scallops data.

Hint: $\hat{\tau}^2$ is the nugget; $\hat{\sigma}^2$ is the partial sill (the sill minus the nugget). Finally, the correlation matrix $H(\hat{\phi})$ can be obtained from the spherical covariance function, part of your solution to Exercise 1.

Note: It appears that `S+SpatialStats` uses the ordinary Euclidean (not geodesic) metric when computing distance, so you may use this as well when computing $H(\widehat{\phi})$. However, you may also wish to experiment with geodesic distances here, perhaps using your solution to Chapter 1, Exercise 7.

CHAPTER 3

Basics of areal data models

We now present a development of exploratory tools and modeling approaches that are customarily applied to data collected for areal units. We have in mind general, possibly irregular geographic units, but of course include the special case of regular grids of cells (pixels). Indeed, the ensuing models have been proposed for regular lattices of points and parameters, and sometimes even for point-referenced data (see Appendix A, Section A.5 on the problem of inverting very large matrices).

In the context of areal units the general inferential issues are the following:

(i) Is there spatial pattern? If so, how strong is it? Intuitively, "spatial pattern" suggests measurements for areal units that are near to each other will tend to take more similar values than those for units far from each other. Though you might "know it when you see it," this notion is evidently vague and in need of quantification. Indeed, with independent measurements for each unit we expect to see *no pattern*, i.e., a completely random arrangement of larger and smaller values. But again, randomness will inevitably produce some patches of similar values.

(ii) Do we want to smooth the data? If so, how much? Suppose, for example, that the measurement for each areal unit is a count, say, a number of cancers. Even if the counts were independent, and perhaps even after population adjustment, there would still be extreme values, as in any sample. Are the observed high counts more elevated than would be expected by chance? If we sought to present a surface of expected counts we might naturally expect that the high values would tend to be pulled down, the low values to be pushed up. This is the notion of smoothing. No smoothing would present a display using simply the observed counts. Maximal smoothing would result in a single common value for all units, clearly excessive. Suitable smoothing would fall somewhere in between, and take the spatial arrangement of the units into account.

Of course, how much smoothing is appropriate is not readily defined. In particular, for model-based smoothers such as we describe below, it is not evident what the extent of smoothing is, or how to control it.