

Nonparametric spatial covariance functions: Estimation and testing

OTTAR N. BJØRNSTAD

NCEAS, 735 State St., Suite 300, Santa Barbara, California 93101-3351

E-mail: bjornsta@nceas.ucsb.edu. Present Address: Department of Entomology, Penn State University, University Park, Pennsylvania 16802

WILHELM FALCK

Department of Biology, Division of Zoology, University of Oslo, Box 1050 Blindern, N-0316 Oslo, Norway

Received March 1999; Revised January 2000

Spatial autocorrelation techniques are commonly used to describe genetic and ecological patterns. To improve statistical inference about spatial covariance, we propose a continuous nonparametric estimator of the covariance function in place of the spatial correlogram. The *spline correlogram* is an adaptation of a recent development in spatial statistics and is a generalization of the commonly used correlogram. We propose a bootstrap algorithm to erect a confidence envelope around the entire covariance function. The meaning of this envelope is discussed. Not all functions that can be drawn inside the envelope are candidate covariance functions, as they may not be positive semidefinite. However, covariance functions that do not fit, are not supported by the data. A direct estimate of the L_0 spatial correlation length with associated confidence interval is offered and its interpretation is discussed. The spline correlogram is found to have high precision when applied to synthetic data. For illustration, the method is applied to electrophoretic data of an alpine grass (*Poa alpina*).

Keywords: bootstrapping dependent data, correlogram, geostatistics, nonparametric regression, population genetics, smoothing spline, spatial autocorrelation

1352-8505 © 2001  Kluwer Academic Publishers

1. Introduction

Spatial autocorrelation techniques are commonly used in population biological inference. It was first introduced in population genetics (Sokal and Oden, 1978a, b), and has recently been used extensively both in genetic and ecological studies (Epperson, 1993a; Epperson and Li, 1997; Bjørnstad *et al.*, 1999a; Koenig, 1999). Theory predicts that the covariance in the genetic makeup of individuals (Lande, 1991) or in the growth of populations (Bjørnstad *et al.*, 1999a) may be a function of the spatial distance separating the sampling units. However, since theory is incomplete, researchers are rarely willing to assume specific functional forms (such as the exponential or the Gaussian). The reason is that quite

1352-8505 © 2001  Kluwer Academic Publishers

diverse forms are possible. The most commonly used method to estimate the relationship between covariance and distance is therefore one that is nonparametric: the spatial correlogram. The spatial correlogram may be interpreted as attaining this by providing an estimate of the spatial autocorrelation function within discretized distance classes. Throughout the text we mean *nonparametric* in the regression sense, i.e., a method that does not assume specific functional forms for the relation. Correlograms have proved very valuable for biological inference. There are, however, two issues for which improvement may be desired: (1) the correlogram approximates the underlying continuous spatial covariance function by a discrete function. The estimator itself is not a valid covariance function, since it is not positive semidefinite. (2) Error bounds for the correlogram are not easily obtained. We introduce and investigate a modified version of the recently developed nonparametric covariance function (NCF; Hall *et al.*, 1994) that may improve on these issues. We call the modification the *spline correlogram*.

In population genetics, spatial covariance denotes the way the genetic composition covaries among individuals distributed through space. Different models predict this covariance to drop according to an exponential, a Bessel or a Gaussian model (e.g., Lande, 1991). Similar predictions can be made in population ecology from simple models for how abundances covary because of movement of individuals (Bjørnstad and Bolker, 2000). These are all functions that can be routinely fitted with geostatistical software (Deutsch and Journel, 1992). Such predictions are extremely simplistic, however. Genetic covariance in plants, for instance, is generated by two different processes that may have unrelated dispersal distance distributions: dispersal of pollen and dispersal of seeds. In the absence of other structuring forces, the final covariance function will be the *convolution* of the seed dispersal distribution with the pollen dispersal distribution (see, e.g., Bjørnstad and Bolker, 2000). The resultant function will often diverge quantitatively from those routinely used in geostatistics. Population ecological theory also offers interesting complications. Spatially extended predator-prey interactions can result in traveling waves in abundance. The covariance function will in such cases be cyclic (Bjørnstad *et al.*, 1999a).

The way the spatial covariance declines with distance is an important probe for testing hypotheses about which processes are involved in shaping spatial patterns in biological populations. This is well illustrated by the mass of studies on genetic microdifferentiation in plants (McGraw, 1995; Linhart and Grant, 1996). Local genetic similarity may arise both as a consequence of adaptation to the local environment (Linhart and Grant, 1996) and as a consequence of constraints on gene dispersal (Epperson and Li, 1997). Patterns of spatial covariance are commonly studied in population biology because different processes may lead to different patterns (Lande, 1991; Epperson, 1993a; Epperson and Li, 1997). Genetic drift (stochastic divergence of local populations) coupled with local dispersal leads to local positive autocorrelation that decrease with distances (Sokal and Wartenberg, 1983; Lande, 1991; Epperson and Li, 1997). Models incorporating these traits are called isolation-by-distance models. Such models predict similar patterns of spatial covariance in different genetic markers sampled from a set of individuals (Sokal and Wartenberg, 1983; Sokal and Jacquez, 1991). The scale of the processes generating microdifferentiation may be reflected in the pattern of spatial covariance (e.g., Epperson and Li, 1997). The “x-intercept” is a frequently employed measure of the scale of pattern in genetic (e.g., Sokal and Wartenberg, 1983; Epperson, 1993a) and ecological studies (Bjørnstad *et al.*, 1999a). Improved methodology to address spatial covariance without assuming *a priori* functional forms will thus help biological inference.

The outline of the paper is as follows: After a very brief preamble pertaining to spatial covariance in population genetics and ecology, we introduce the nonparametric covariance function (NCF) and the related spline correlogram and describe its relationship to the traditional spatial correlogram. Then we describe a bootstrap algorithm to provide a confidence envelope for the entire estimator and its derived statistics. We investigate through Monte Carlo simulations the precision of the spline correlogram and the coverage of the bootstrap confidence envelope. We further discuss how the method can be extended to vectorial data (e.g., multiple genetic markers or time series). We apply the method to synthetic data with known spatial covariance. For illustration we also analyze genetic data on an alpine meadow grass, *Poa alpina* L.

2. Methods and analyses

2.1 Preamble

Because of the theoretically diverse predictions about functional forms (see Introduction), this paper will concern itself with nonparametric estimators. That is not to belittle the large and important theory pertaining to parametric covariance functions as developed in geostatistics (e.g., Deutsch and Journel, 1992; Cressie, 1993). It is rather a reflection of the goals being different; in geostatistics modeling the covariance is typically just a small step towards the final objective of spatial prediction and interpolation. In population biology, in contrast, the covariance function is—as detailed in the introduction—a goal in itself because it provides a bridge between the theory of biological processes and data. In this respect, the most relevant result from geostatistical theory pertains to permissible models for covariance (e.g., Zimmerman, 1989): To qualify, a potential function has to be positive semidefinite. We return to this below. In the development on the methodology and in the discussion we will, throughout, assume that data to come from a second order stationary field (Cressie, 1993). That is, we assume the expectation, and the covariance function to be unchanging through space. We will further assume the field to be isotropic, so that the covariance only depends on distance and not direction. Unless otherwise stated we assume the data to be Gaussian. However, in the cases where we discuss the covariance in the genotype of individuals (see also Epperson, 1995), we work with spatially distributed categorical variables. The notion of covariance functions is more complicated in such systems. We implicitly treat these as *hidden* Markov random fields (e.g., MacDonald and Zucchini, 1997), realized according to a binomial or multinomial filter. The covariance we estimate is thus that of the underlying field (see also Albert and McShane, 1995). Whether this is a useful heuristic for specific genetic systems is a discussion outside the scope of the current study.

2.2 Nonparametric covariance functions: the spline correlogram

Consider the measurement z_i on an individual i at coordinate $\{x_i, y_i\}$. Assume that there are n individuals, and that the pairwise spatial covariance is a function, $\rho(\delta)$, of the distance, δ , separating the individuals. The geographic distance, δ_{ij} , between individuals i and j is:

$$\delta_{ji} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \quad (1)$$

The sample autocovariance between the two is:

$$\text{Cov}(z_i, z_j) = (z_i - \bar{z})(z_j - \bar{z}), \quad (2)$$

where $\bar{z} = 1/n \sum_{l=1}^n z_l$ is the sample mean. (Note that this measure is a consistent estimator of the population mean, but it may for a given sampling design be biased in the presence of spatial dependence). The sample autocorrelation between individuals i and j is then estimated as:

$$\hat{\rho}_{ij} = \hat{\rho}(z_i, z_j) = \frac{(z_i - \bar{z})(z_j - \bar{z})}{1/n \sum_{l=1}^n (z_l - \bar{z})^2}. \quad (3)$$

Among n individuals there are $n(n-1)/2$ unique pairwise autocorrelations, corresponding to the upper (or lower) triangle of the sample autocorrelation matrix (that is, $\hat{\rho}_{ij}$ for $i = 1, \dots, n, j = i+1, \dots, n$).

Hall and Patil's (1994, Equation (2.1)) kernel estimator of autocorrelation as a function of geographic distance is given by:

$$\tilde{\rho}(\delta) = \frac{\sum_{i=1}^n \sum_{j=1}^n K(\delta_{ij}/h) (\hat{\rho}_{ij})}{\sum_{i=1}^n \sum_{j=1}^n K(\delta_{ij}/h)}, \quad (4)$$

where K is a *kernel function* (e.g., Härdle, 1990) and $h(>0)$ is the *bandwidth*. The bandwidth is the parameter that adjusts the smoothness of the fitted curve. This parameter is analogous to the distance-class width in the spatial correlogram (see below). The estimate $\tilde{\rho}(\delta)$ will be a function that is nonparametric in the sense of not assuming any specific class of parametric models for the relation (e.g., Härdle, 1990). Hall and Patil (1994) proved that the kernel estimator (4) can be tuned (by tuning h) so that $\tilde{\rho}(\delta) \rightarrow \rho(\delta)$ as $n \rightarrow \infty$ for *any* smooth functional form of $\rho(\delta)$. That is, as long as the true covariance function is C_2 -differentiable (has continuous 1st and 2nd derivatives), the kernel estimator is a consistent estimator. Note, though, that discontinuity of the derivatives may be permitted at the boundary $\delta = 0$. We use a cubic B-spline as an equivalent kernel smoother (Nychka, 1995) because this adapts better to irregularly spaced data than many regression kernels (see, for example, Jones *et al.*, 1994) and is known to provide consistent estimates of the covariance function (Hyndman and Wand, 1997). The asymptotic kernel function for the cubic B-spline (Green and Silverman, 1994) is:

$$K(u) = \frac{1}{2} \exp\left(-\frac{|u|}{\sqrt{2}}\right) \sin\left(-\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right), \quad (5)$$

where u is the argument: i.e., δ/h in our case. This asymptotic kernel (5) is nearly bell-shaped (except for having sinusoidal tails that vanish rapidly). Note, that we use standard techniques for fitting the spline function (e.g., Green and Silverman, 1994), and that the asymptotic kernel (5) is presented to clarify the link between the NCF and the spline correlogram.

A standardized way to report the degree of smoothing of a spline fit is to calculate the *equivalent degrees of freedom* (edf) (Green and Silverman, 1994). The equivalent df measures the number of effective parameters (in a multiple regression or polynomial regression sense) that are used in the fit. This number is defined as the trace of $(\mathbf{I} - \mathbf{S})$,

where \mathbf{I} is the identity matrix and \mathbf{S} is the *smoother matrix* associated with the spline fit (e.g., Green and Silverman, 1994: chapter 3). In the analyses below we report the edf rather than the bandwidth unless otherwise stated.

The above discussion ensures pointwise consistency of $\tilde{\rho}(\delta)$. The estimator may still violate the basic requirement of positive semidefiniteness, however. Positive semidefiniteness implies (by Bochner's theorem) that the Fourier transformed function is strictly nonnegative (Hall *et al.*, 1994). This is not guaranteed by Equation (4). We use Hall *et al.*, (1994) Fourier-filter method to ensure positive semidefiniteness. We first obtain the Fourier transform $\tilde{\rho}$. Prior to back-transformation, we ensure nonnegativity by setting all negative excursions of the transformed function to zero. We call the resultant nonparametric estimate of the spatial covariance the *spline correlogram*.

To illustrate the method, we apply it to synthetic data with different but known functional forms for the covariance. We let the covariance, ρ_{ij} , between the random variables z_i and z_j be a function of the distance, δ_{ij} , separating the two. We assume that the covariance follows one of three functional forms in the synthetic data. The exponential (1st order autoregressive) (Fig. 1B)

$$\rho(\delta) = \exp(-\delta/a), \quad (6)$$

where, the parameter a controls the rate of decay in covariance with distance. The Gaussian (Fig. 1C)

$$\rho(\delta) = \exp(-\delta^2/a^2); \quad (7)$$

And the second order spatial process (Fig. 1D and Fig. 1 legend). Note that the ‘‘second order’’ in this context refers to a spatial process that is governed by dependence at two different spatial scales. (It is not to be confused with the order of stationarity of the field—it is still second order stationary).

Data from spatial maps with prespecified spatial covariance structure were generated by drawing (without replacement) n random locations on a 100-by-100 grid. A target covariance matrix A between the n units (individuals) was generated by calculating the distances between the locations and thereafter evaluating the covariance function (e.g., Equation (1)) in each given pairwise distance. A vector, \mathbf{Z} , of multivariate normal data with the target covariance was subsequently generated using the eigendecomposition method (see Ripley, 1987, for details). A mean of zero and a variance of one was used in all cases. Fig. 1 depicts surfaces from realizations of multivariate random processes with the three types of covariance structures. Data on the genotype of individuals is usually categorical, representing presence or absence of electrophoretic bands or DNA bases. We therefore also converted the synthetic data into multivariate binomial data (through a sign transformation; setting negative values to zero and positive values to one).

Spline correlograms for data with the different covariance structures are shown in Fig. 1. The spline correlogram generally adapts well to the different underlying covariance structures. The underlying covariance function is also recovered with reasonable accuracy when the data are binary (Fig. 1 insets). Complete spatial randomness generates a flat covariance function that is centered on zero (Fig. 1A).

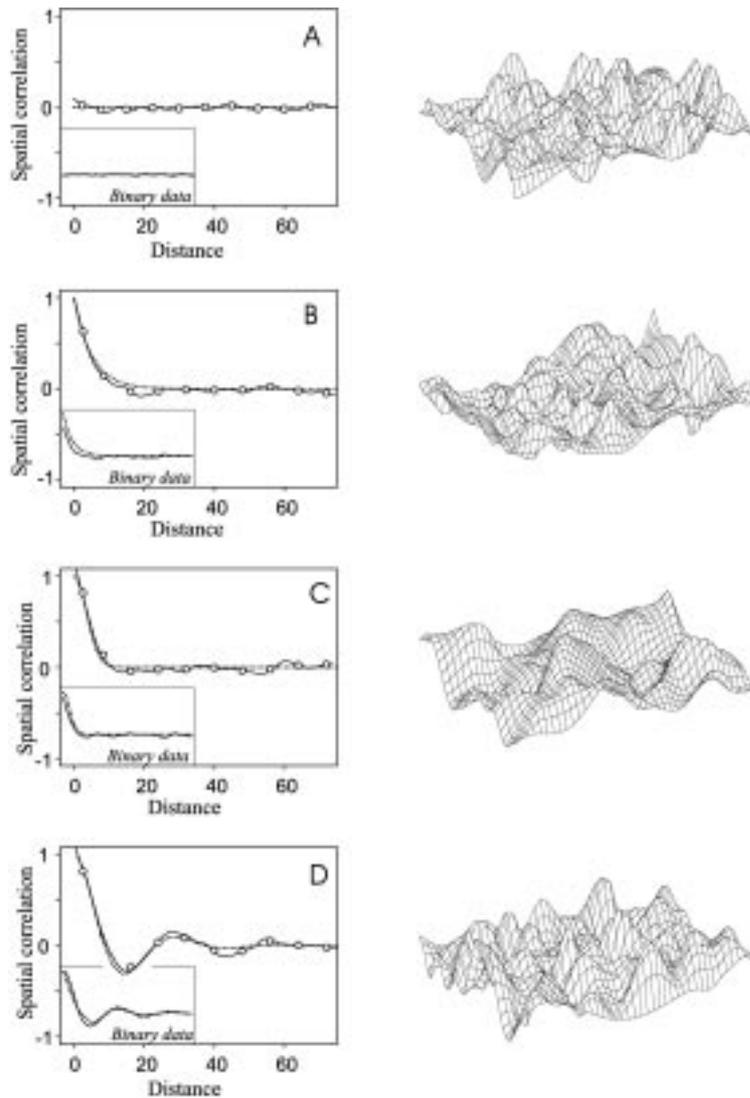


Figure 1. Spatial correlation function (left panels) of four synthetic data sets with known covariance (dotted lines in right panels). The sample size is 400 in each case. The surfaces are interpolations between the 400 observations to visualize the autocorrelated random variables. The spatial correlation is estimated using a spline correlogram (full line) with 40 equivalent degrees-of-freedom, and the spatial correlogram (open circles) with class width of 8. Estimates based on binary data are presented as insets. See text for details on the methods. A. Independent data (no spatial covariance). B. Data from a map with exponential covariance (Equation (6)) with parameter $a = 5$. C. Data from a map with Gaussian covariance (Equation (7)) with parameter $a = 5$. D. Data from a map with 2nd order spatial covariance (see text) with parameters $a_1 = 1.8$ and $a_2 = -0.85$ for which the theoretical covariance function is $\rho(\delta|a_1 = 1.8, a_2 = -0.9) = (-a_2)^{\delta/2} \sin(e\delta + F) / \sin(F)$ where, $e = \arccos(1a_1/2\sqrt{-a_2})$, $F = \arctan[\tan(e)(1 - a_2)/(1 + a_2)]$. See, for example Box and Jenkins (1970) and Cressie (1993) for details on the covariance functions.

2.3 The spatial correlogram

The traditional nonparametric method to measure how covariance is a function of spatial distance uses the spatial correlogram, $C(d)$. It is therefore of interest to consider how $C(d)$ relates to $\tilde{\rho}(\delta)$. The spatial correlogram attains a model-free fit of $\rho(\delta)$ as a function of δ , by quantifying the spatial covariance function at a set of discrete focal distances. In order to see the link between the spatial correlogram and the spline correlogram, we will use a slightly unusual interpretation. We will consider the spatial correlogram as a k -step function, $C_k(d)$, approximating the underlying continuous function $\rho(\delta)$. The traditional spatial correlogram based on Moran's I is then the step function obtained through local averaging of $\hat{\rho}_{ij}$ (as defined in Equation (3)) around the k focal distance d_1, \dots, d_k (Cressie, 1993, chapter 2.4):

$$C_k(d_k | k = 1, \dots, k_{\max}) = \frac{1}{|I_k(\delta)|} \sum_{i_k(\delta)} \frac{(z_i - \bar{z})(z_j - \bar{z})}{1/n \sum_{i=1}^n (z_i - \bar{z})^2} = \frac{1}{|I_k(\delta)|} \sum_{I_k(\delta)} \hat{\rho}_{ij}, \quad (8a)$$

where $I_k(\delta)$ indicates all pairs $\{i, j\}$ for which the geographic distance, δ_{ij} , is within a tolerance region of d_k , and $|I_k(\delta)|$ is the number of distinct pairs in $I_k(\delta)$. The spatial correlogram can thus be written as a sequence of local averages of $\hat{\rho}_{ij}$:

$$C_k(d_k) = \text{mean}_k(\hat{\rho}_{ij} | L_k < \delta_{ij} \leq U_k), \quad (8b)$$

where L_k and U_k signify the lower and upper tolerance limits (usually set such that each pairwise similarity is only used once) around the focal distances. The distance class width, given by $\lambda_k = U_k - L_k$, controls the resolution and complexity of the k -step function. Note, that this interpretation of the spatial correlogram as a step function for the covariance is a heuristic; Equation (8) is not itself a covariance function since it is not positive semidefinite.

There is a dependence of the shape of the correlogram on the centering of the focal distances; Choosing $L_1 = 0$, produces a different result than centering the first bin on 0 (i.e., $L_1 = -\lambda_1/2$) (see Scott, 1992, chapter 5). One may thus consider a sequence of different correlograms for a data set, each differing in the location of the target distances, d_k 's, and then average across these. Assume that m such correlograms are calculated by fixing L_1 between 0 and $\lambda - 1/m$ (i.e., $L_1 = \{s\lambda/m\}$, where $s = 0, \dots, m-1$). The distance classes of each step function will partly overlap, so that an "average shifted" correlogram (cf. Scott, 1992, chapter 5.3) may be obtained as the average across the sequence according to:

$$\bar{C}_{k'} = 1/m \sum_{i=1}^m C_k. \quad (9)$$

If each individual correlogram is evaluated at k locations, then the average shifted correlogram will be evaluated at $k' = km$ locations (m locations within each original distance class). For infinitesimally small shifts, Equation (9) is equivalent to (4) because:

$$\lim_{m \rightarrow \infty} \bar{C}_{k'}(d_{k'}) \rightarrow \tilde{\rho}(\delta), \quad (10)$$

where $\tilde{\rho}(\delta)$ is the particular kernel estimator with a triangular kernel of bandwidth λ (Scott, 1992). There is hence an asymptotic equivalence between the spline correlogram and the spatial correlogram.

2.4 Estimation uncertainty

The sampling variability of the spline correlogram can be found by Monte Carlo simulations (e.g., Manly, 1997) on data for which the true covariance function is known. That is, repeated estimation from a large number of synthetic data sets all with the same covariance structure. Fig. 2 shows the variability of the spline correlogram for exponential and Gaussian covariance functions with different parameters. The variability is calculated from 1000 simulated data sets (generated as outlined above) with 250 observations in each. The estimated spline correlograms are usually close to the true covariance function. The mean correlation between the estimate and the true function are generally around 0.95 for these data sets (exponential $a = 5$: mean = 0.94, sd = 0.03; exponential $a = 10$: mean = 0.94, sd = 0.03; Gaussian $a = 5$: mean = 0.96, sd = 0.02; Gaussian $a = 10$: mean = 0.95, sd = 0.02). The results indicate relatively high precision of the

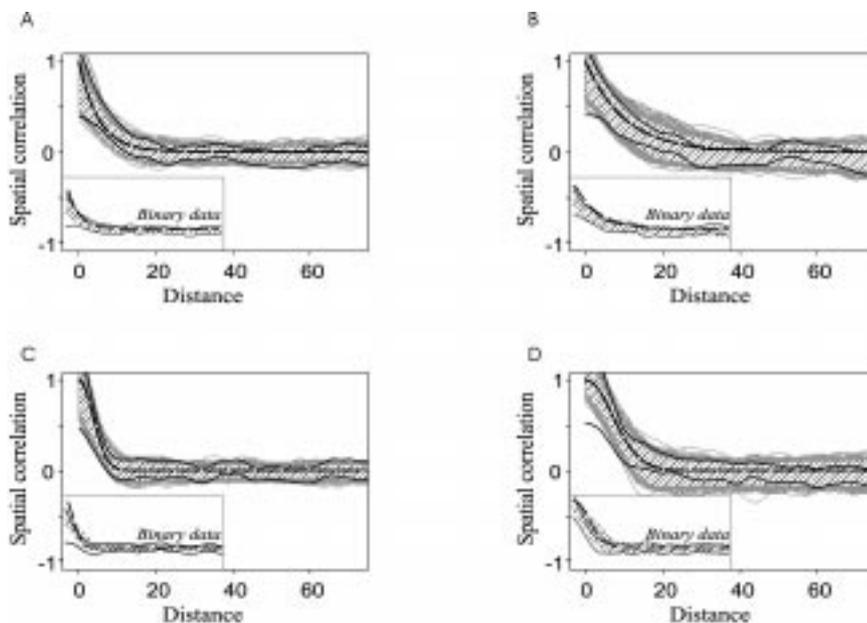


Figure 2. Spline correlograms (25 edf) was applied to 1000 data sets with one of four covariance structures to quantify the estimation uncertainty. The sample size was 250 in each case. The hatched regions represent the empirical 95% distribution based on estimates for 1000 data sets for each covariance function. The light shaded area represents all the 1000 estimates. The broken lines represent the true covariance functions. Thick lines represent the estimate and the 95% confidence envelopes on the basis of bootstrapping from a single data set (using 1000 bootstrap resamples). A and B are data from maps with exponential covariance (see Equation (6); A with parameter $a = 5$ and B with parameter $a = 10$). C and D are data from maps with Gaussian covariance (see Equation (7); C with parameter $a = 5$ and D with parameter $a = 10$). Estimates based on binary data are presented as small insets.

method in the sense that the estimated functions are closely correlated to the true functions. The precision is marginally reduced when the data are binary (insets in Fig. 2). The estimated spline correlograms estimated from the binary data are still close to the true covariance function. The mean correlation between estimate and true function is a little below 0.95 (exponential $a = 5$: mean = 0.92, sd = 0.04; exponential $a = 10$: mean = 0.93, sd = 0.04; Gaussian $a = 5$: mean = 0.95, sd = 0.03; Gaussian $a = 10$: mean = 0.95, sd = 0.02). Note, though, that there is a negative bias of the local autocorrelation when the data is binary (the estimates are biased towards zero, for local distances). This effect is previously documented and was discussed by Epperson (1995) for the spatial correlogram.

2.5 Confidence regions

For real data it is necessary to obtain a measure of estimation uncertainty from a single realization of the sampling process. We use a dedicated bootstrap algorithm to generate a confidence envelope for the nonparametric covariance function. We will discuss the interpretation of the confidence envelope in the discussion. The bootstrap algorithm proceeds as follows (see also Efron and Tibshirani, 1993): first, bootstrap data sets are generated by sampling (with replacement) from the individual observations—that is, sampling among the tuples consisting of the spatial coordinates (x and y) and the genetic score (z). We draw n observations from the original n observations $\{x, y, z\}_i$, where $i = 1, \dots, n$. The spline correlogram is subsequently calculated from the bootstrap data set. Pairs of distances $\{\delta_{ii}, \hat{\rho}_{ii}\}$ between one individual with itself (due to the sampling with replacement) are discarded prior to calculations to avoid bias for short distances. The whole process is repeated to give a bootstrap sampling distribution for the spatial covariance function. A confidence envelope is erected by the quantile method (Efron and Tibshirani, 1993), where the $\alpha\%$ -level confidence envelope of the estimator is given by the $\alpha/2\%$ and $(100 - \alpha/2)\%$ quantiles of the bootstrap distribution.

Since the spline correlogram is a relatively complicated estimator, we need to study how the bootstrap performs numerically (Young, 1994). This is a large task to do properly. Here we report on a preliminary study. We thus assessed the success of the bootstrap through Monte Carlo simulations by comparing confidence envelopes estimated by bootstrapping single realizations with the sampling variability found by Monte Carlo methods (Fig. 2). The typical 95% confidence interval should correspond to the 95% estimation variability of the estimator (e.g., Efron and Tibshirani, 1993). Confidence envelopes from bootstrapping single data sets are shown in Fig. 2. The envelopes generally relate well to the sampling variability of the estimator. The correlation between the 95% confidence interval and the 95% sampling distribution is generally higher than 0.9 for the type of synthetic data depicted in Fig. 1 (exponential $a = 5$: correlation = 0.9; exponential $a = 10$: correlation = 0.93; Gaussian $a = 5$: correlation = 0.95; Gaussian $a = 10$: correlation = 0.98). At a first glance, bootstrapping thus appears to give a reasonable estimate of the sampling variability. The correspondence is also good for the binary data. The performance is, however, reduced somewhat due to the nominal nature of the data. Hall *et al.* (1994; see also Hall, 1993) also discusses the use of the bootstrap for the NCF.

2.6 The correlation length

The L_0 correlation length is calculated from the spline correlogram as the smallest value for δ such that $\tilde{\rho}(\delta) = 0$. We investigate this measure because it is frequently used in population biology. We will briefly discuss its meaning (if any) in the discussion. To investigate this statistic we resort to caricatured binary (0/1) maps with idealized patch structure (Bjørnstad and Falck, 1997). A pattern of constant sized patches (with radius r) separated by r units was generated to form a checkerboard. In this way any two randomly chosen points that are less than the distance r away from each other are likely to have the same value. Thus, the correlation length is r in these data sets. (Note that these maps are not meant to have arisen from any population biological process. The maps are designed to give a clearly defined correlation length.) A random sample of 250 points was selected from the maps. For the data sampled from patchy binary maps ($n = 250$) with different patch configuration (patch radius: 2, 5, 10, 15 and 25), the point estimates agree well with the true correlation length (Table 1). An exception is in the case where the patch size is small relative to the grain (resolution) of the sampling grid. The correspondence between the empirical sampling variability and the bootstrap confidence interval is satisfactory (Table 1).

In summary, these preliminary analyses of synthetic data indicate that the spline correlogram is capable of recovering functionally very different covariance structures. The bootstrap appear to give reasonable estimates of the sampling variability and appears to provide confidence envelopes with appropriate coverage. For illustration we now analyze a real set of spatial genetic data.

2.7 Data on *Poa alpina*

This data set is based on previously published ecological and genetic data (Nordal and Iversen, 1993; Bjørnstad *et al.*, 1995). Two-hundred-and-forty-nine individuals were sampled within 15 quadrats (10×10 m). The quadrats were unevenly distributed along

Table 1. The estimated L_0 correlation lengths (x -intercept) estimated from binary maps with known correlation lengths. The estimates are based on spline correlograms with 25 equivalent df applied to data sets with 250 observations. The results are given for 5 different patch radii (r). The interpatch distance is set to r in all cases so as to give maps with theoretical L_0 correlation lengths equal to r . Monte Carlo summarizes the mean (2.5- and 97.5-percentiles) estimates across 1000 Monte Carlo runs. Bootstrap 1–3 report the estimate and 95% bootstrap confidence intervals (1000 bootstrap resamples) estimated from individuals realizations of each configuration.

	Monte Carlo	Bootstrap 1	Bootstrap 2	Bootstrap 3
$r = 2$	3.2 (0, 17.4)	13.9 (0, 25.1)	9.5 (0, 19.7)	0 (0, 23.2)
$r = 5$	5.5 (4.8, 6.3)	4.8 (3.3, 6.2)	5.1 (4, 6.1)	5.2 (4.3, 6.1)
$r = 10$	10.8 (9.8, 11.9)	11 (9.6, 12.3)	11.6 (10.4, 12.9)	10.9 (9.5, 12.4)
$r = 15$	16.6 (15, 18.6)	17.8 (15.5, 20.6)	16.6 (15.1, 18.4)	17.6 (15.1, 19.9)
$r = 25$	27.7 (23.4, 32.7)	30.2 (22.1, 36)	28.5 (24.3, 31.5)	28.3 (25.5, 31.9)

three transects, each covering a gradient from sheltered snow beds to exposed ridge habitat. The median distance between neighboring quadrats is 30 m (range: {10 m, 90 m}). The three transects had roughly the same bearing and were 250 m, 800 m, and 1200 m apart, respectively. The entire study area extends across 1.5 km of alpine vegetation. Within each quadrat, the exact location of individuals was not recorded. The individual may nevertheless be used as the statistical unit (Epperson, 1995). The distance between individuals within the same plot is arbitrarily set to zero, so that zero-distance covariance represented within-quadrat similarity. Each individual was screened for genetic composition at 4 polymorphic isozyme loci (three of which had one variable band and one with three variable bands giving a total of 6 variable bands) using enzyme electrophoresis (see Nordal and Iversen, 1993, for details). Several aspects of the relationship between genetic similarity and spatial distance are of particular interest: (1) Evidence of microdifferentiation through significantly positive autocorrelation of individuals close in space; (2) The L_0 correlation length; (3) The overall shape of the covariance function; (4) The uncertainty associated with the estimated covariances; and (5) Whether there are significant differences between different isozyme systems.

Spatial covariance functions were estimated for the 6 polymorphic alleles using the spline correlogram (with 25 edf and 1000 bootstrap resamples). The within-quadrat autocorrelation, $\hat{\rho}(0)$, and the L_0 correlation length are tabulated in Table 2. The within-quadrat autocorrelation is significantly positive for all systems, indicating non-random local structuring. The within-quadrat autocorrelation is not significantly different for the six systems at a Bonferroni-corrected 5% level (Table 2). The estimated correlation length range from 20–100 m. Most systems have similar estimates. However, SKD1 has a shorter correlation length than the other allele systems. The difference is statistically significant at a Bonferroni corrected 5% level against 2 (SKD67, SKD100) of the other 5 (Table 2). The spline correlograms (Fig. 3) are more complicated than predicted by simple theoretical models. Although, all systems exhibit a covariance that drops with distance initially, there is evidence of significant negative autocorrelation in four of them (Fig. 3). Superimposing the confidence envelopes of the covariance function reveal no significant differences between the spatial covariances of most systems. The spatial covariance function of SKD1, however, drops significantly (at the 5% level) more quickly than that of all the other systems (Fig. 3).

Table 2. Estimates based on spline correlograms (25 equivalent df) of the 6 allele systems of *Poa alpina*. The x -intercept, L_0 , is the estimate of the distance at which the genetic similarity of two individuals is no different that expected by chance alone within the sample. The estimated within-quadrat similarity is given by $\hat{\rho}(0)$. The 95% confidence intervals (CI) are estimated using bootstrapping based on 1000 bootstrap samples. Letters in bold (A and B) identify groups of estimates that are significantly different at a Bonferroni-corrected 5%-level.

	<i>GPI</i>	<i>SKD1</i>	<i>SKD67</i>	<i>SKD100</i>	<i>MDH</i>	<i>PGM</i>
L_0	108.6 (18.1, 131.3)	20.0 (16.7, 23.5)	65.0 (55.0, 82.3)	88.9 (32.8, 116.8)	50.4 (16.8, 67.8)	37.3 (21.5, 81.6)
$\hat{\rho}(0)$	0.26 (0.14, 0.42)	0.55 (0.41, 0.71)	0.42 (0.24, 0.68)	0.47 (0.34, 0.64)	0.18 (0.08, 0.31)	0.44 (0.20, 0.80)

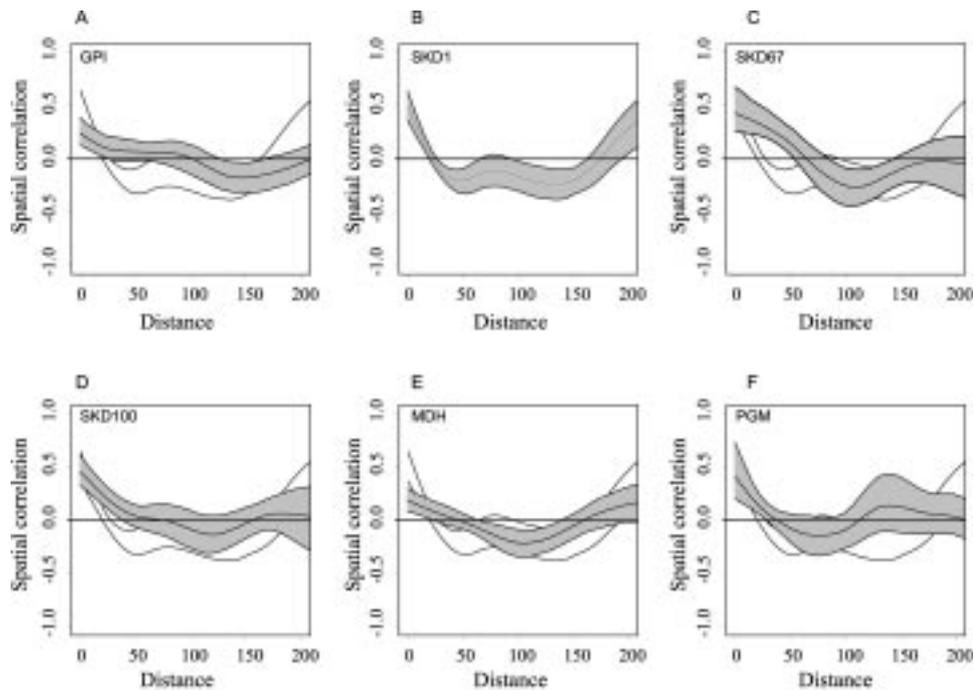


Figure 3. Spline correlograms (25 edf) with 95% bootstrap confidence envelope for the covariance function (shaded regions) of the 6 polymorphic isozyme bands of *P. alpina*. Significant positive spatial autocorrelation at short distances is evident for all systems. Several systems exhibit significantly negative autocorrelation for distant individuals. The confidence envelope of SKD1 (the two black lines represent the upper and the lower confidence envelope) are superimposed on the other 5 correlograms illustrating how SKD1 has a spatial covariance profile that drops significantly more quickly than all the other allele systems.

3. Discussion

In this paper we introduced the *spline correlogram* to estimate the spatial covariance function nonparametrically. The motivation for a nonparametric estimator is two-fold. Firstly, current population biological theory suggests that plausible spatial processes may give rise to functionally diverse patterns of covariance. The class of plausible functional forms is wider than that commonly used in geostatistics. Secondly, the nonparametric estimator may provide a complementary way of assessing how closely data match our preconceptions about functional forms. A confidence envelope for the entire covariance function is calculated through bootstrapping. Through the spline correlogram we obtain a direct estimate of the spatial correlation length. In the following discussion we first focus on the technical issues and subsequently briefly discuss the genetic inference with respect to the sample data.

3.1 Confidence envelope

Our Monte Carlo study of the bootstrap estimator of the confidence envelope is preliminary. The numerical simulations suggest, nevertheless, that the envelopes have the appropriate coverage. Ultimately, a much more extensive investigation must be undertaken. Our main purpose here is to introduce the main idea. The meaning of the envelope deserves some discussion. In order to qualify as a covariance function, a candidate has to be positive semidefinite. It will, however, always be possible to draw curves inside the envelope that fail to fulfill this. So not all functions that fit inside the envelope are permissible functions. A more important property for inference, however, is that candidate covariance functions that do not fit inside the envelope, are not supported by with the data. In this way, the envelopes provides a method for model checking (see Tsay, 1992, for a related discussion). A corollary is that data with non-overlapping confidence envelopes must come from random fields with different underlying covariance structure.

By construction, the confidence envelope is a point-wise confidence interval. The curvewise error rate is therefore likely to be larger than the nominal 5% level. The Bonferroni correction has been used to control the correlogram-wise error rate of the traditional correlogram (Oden, 1984). This may be advised against in the spline correlogram because the class of Bonferroni corrections fails to make use of the large positive correlation between nearby points of the curve (Härdle, 1990: chapter 4). For the present we resort to treating the significance level of any contrast with caution.

3.2 Correlation length

Different measures and interpretations of the spatial correlation length can be found in the literature. In population biology, it has been common to use the x -intercept—that is, to use $\rho(\delta) = 0$ as the reference line—and estimate the distance $\hat{\delta}_0$ such that $E[\rho|\delta = \hat{\delta}_0] = 0$ (Sokal and Wartenberg, 1983; Epperson, 1993a). To use the zero-covariance as the reference line is however not trivial, since many theoretical covariance functions (e.g., the exponential and the Gaussian) tend only asymptotically to zero. With some hesitation we therefore employ the slightly modified definition: the distance at which the covariance is not *significantly* different from 0. A common alternative in other disciplines is the $L_{1/e}$ correlation length (Myers *et al.*, 1995; Hilgers *et al.*, 1996), for which $\rho(z_i, z_j) = e^{-1} \approx 0.37$ represents the reference line. $L_{1/e}$ will be an estimate of the fundamental parameter if the underlying covariance function is exponential (Equation (6)). A disadvantage of this reference point is that it may not have any particular interpretation for other covariance functions. One useful interpretation of the x -intercept arise when the sample average is used to center the covariances (as in Equation (2)). In such a case, the sample L_0 correlation length is an estimate of the distance across which two measurements is no more similar than the average sample similarity. It should be stressed that, as for all statistical parameters, there may be a bias between the sample measure and the true population property. The “average similarity” is a function of the sample mean (cf. Equations (2) and (3)). The sample correlation length is therefore a function of the sample average, and will thus depend on the scaling of the study relative to the correlation length in the field. The dependence is weak when the extent (size of the study area) is reasonably

large relative to the correlation length of the processes (because the sample average quickly converges on the population mean). However, for long correlation lengths, there will be a negative bias in the estimate that can be severe (sometimes called the ‘‘volume effect’’, Bayly *et al.*, 1993). A second bias occurs when the correlation length is short relative to the grain (resolution) of the study (see Table 1). As long as the spatial design is the same, biases will be similar so that contrasts between different systems should be relatively robust. Comparisons of spatial covariance of different loci screened from the same individuals should therefore be meaningful (see Bjørnstad *et al.*, 1999b, for an example from population ecology).

3.3 A multivariate covariance function

Multivariate genetic data is increasingly common. DNA sequencing generates series of variables in the form of presence or absence of base pairs (Bertorelle and Barbujani, 1995). Also isozyme data may require multivariate techniques. When microdifferentiation is due to drift and local dispersal, for instance, the spatial covariance will be similar for all loci (Sokal and Wartenberg, 1983; Sokal and Jacquez, 1991; Epperson, 1993b). Increased accuracy may be gained through a multivariate covariance function that summarizes the pattern across the different descriptors. The classical multivariate method to estimate spatial covariance is the Mantel correlogram (Oden and Sokal, 1986; Legendre and Fortin, 1989). The Mantel correlogram $R(d)$ corresponds to a spatial correlogram (cf. Equation (8a)), in which the observation z is vectorial. Replacing the scalar product by the vector product in Equations (4) and (8) we may define:

$$R_{ij} = R_{ji} = \tilde{z}_i^T \tilde{z}_j, \quad (11)$$

where T denotes matrix transposition, and \tilde{z} is the matrix where each column (allozyme band; $m = 1, \dots, M$), z_m has been rescaled (normalized) according to

$$\tilde{z}_m = \frac{z_m - \bar{z}_m}{\sqrt{\sum_{l=1}^n (z_l - \bar{z}_m)^2 / n}}.$$

The Mantel correlogram may be written as the sequence of local averages of R_{ij} against the Euclidean spatial distance (as in Equation (8b)). Thus, replacing the sample covariance $\hat{\rho}_{ij}$, with the multivariate analogue, R_{ij} , the multivariate covariance function may be estimated as:

$$\tilde{R}(\delta) = \frac{\sum_{i=1}^n \sum_{j=1}^n K(\delta_{ij}/h) (R_{ij} - \bar{R})}{\sum_{i=1}^n \sum_{j=1}^n K(\delta_{ij}/h)}, \quad (12)$$

where R_{ij} is defined in (11), \bar{R} is the sample average of the R_{ij} 's and all other symbols and functions are as in (6). Bjørnstad *et al.* (1999b) illustrates the use of the multivariate spline correlogram to estimate spatial covariance in ecological data. We provide an example for genetic data below. A possible extension of the estimator (12) for multivariate data is that obtained through centering the components in (11) row-wise. The resultant model may prove to be interesting as an estimator of a generalized covariance function (i.e., one in

which the spatial expectation is allowed to vary between locations). A discussion of this is however outside the scope of the current discussion.

3.4 Model complexity

A choice has to be made with respect to complexity of the model used to estimate the spatial covariance function. This complexity is controlled by the width of the distance classes in the spatial correlogram, and by the bandwidth (or equivalent degrees-of-freedom) of the spline correlogram. The choice of complexity is nontrivial since the data are non-independent. This topic will require future study. Currently we will just point out an interesting level of robustness of the spline correlogram relative to the spatial correlogram. For the spatial correlogram, narrow distance classes give high resolution but low precision. Wide distance classes give stable estimates but low resolution. The nonparametric covariance function, in contrast, appears relatively robust with respect to the complexity because the bandwidth and local resolution are less closely linked. Unpublished numerical experiments (Bjørnstad and Falck, 1997) testify that overly restrictive models (10 df or less) appear biased. The x -intercept, for example, overestimated the correlation length and gave unstable inference for a number of spatial configurations (see also Table 1); A range of more complex models, however, appear to exhibit low bias. Of course, excessively complicated models tend to have inflated variance.

3.5 Genetic inference for *Poa alpina*

Applying the spline correlogram to the data set of *P. alpina* give correlation lengths around 50 m (varying between 25 and 100 m). The multivariate spline correlogram for the allozyme bands of *P. alpina* (except SKD1) estimate the L_0 correlation length at 57.4 m {48.7, 69.2}. The spatial covariance functions of one of the six variable isozyme bands (SKD1) drops significantly faster than the others do. The remaining five bands have covariance functions that decrease at similar rates with distance. A process of isolation-by-distance may be seen as consistent with such similarity. Because the reproduction of *P. alpina* is partly asexual (apomictic), this pattern of spatial covariance may reflect the bulbil ("seed") dispersal distances (Bjørnstad *et al.*, 1995). The divergence in SKD1, and the presence of negative autocorrelation in several of the isozyme bands is, however, contrary to an isolation-by-distance model (Sokal and Jacquez, 1991; Epperson, 1993a). From a theoretical point of view, such patterns of covariation are expected from clinal selection. The scale of the dissimilarity corresponds loosely to the length of the three transects spanning the snowbed-ridge gradient (above). These environmental features may thus offer selection gradients of the right type. Bjørnstad *et al.* (1995) discusses the system in more detail.

3.6 Nonparametric variography

Beside the spatial correlogram, the variogram is a frequently used method to quantify spatial pattern (Cressie, 1993). The scaled variogram may be represented as a step function

(like the correlogram) that locally averages the Geary distance (a scaled Euclidean distance) instead of the sample autocorrelation (in the fashion outlined in Equation (8b)). Since the variogram has certain advantages over the correlogram (e.g., Cressie, 1993), it may be worthwhile to investigate the possibility of extending the continuous nonparametric regression setting to such an estimator. Technically speaking, such an extension is easily implemented by replacing $\hat{\rho}_{ij}$ with the Geary distance in Equation (4). Future theoretical developments may show whether this is a fruitful line of inquiry.

A main strength of the spline correlogram is its use to draw statistical inference about differences and similarities in the spatial covariance of different genetic or ecological systems. The bootstrap confidence regions allow formal testing of differences among different measures. Through this we believe we open for a more rigorous testing of the expanding body of theory pertaining to genetic and ecological differentiation.

Acknowledgment

We want to thank Bryan K. Epperson, Rolf A. Ims, Pierre Legendre, Nils Chr. Stenseth, Nigel G. Yoccoz and two anonymous reviewers for discussion and comments on statistics and biology. Funding was received from the National Center for Ecological Analysis and Synthesis (a Center funded by NSF Grant DEB-94-21535, the University of California Santa Barbara, and the State of California) and from the Norwegian National Science Foundation (ONB).

References

- Albert, P.S. and McShane, L.M. (1995) A generalized estimating equations approach for spatially correlated binary data: Applications to the analysis of neuroimaging data. *Biometrics*, **51**, 627–38.
- Bayly, P.V., Johnson, E.E., Wolf, P.D., Greenside, H.D., Smith, W.M., and Ideker, E.E. (1993) A quantitative measurement of spatial order in ventricular fibrillation. *Journal of Cardiovascular Electrophysiology*, **4**, 533–46
- Bertorelle, G. and Barbujani, G. (1995) Analysis of DNA diversity by spatial autocorrelation. *Genetics*, **140**, 811–19.
- Bjørnstad, O.N., Iversen, A., and Hansen, M. (1995) The spatial structure of the gene pool of a viviparous population of *Poa alpina*—environmental controls and spatial constraints. *Nordic Journal of Botany*, **15**, 347–54.
- Bjørnstad, O.N. and Falck, W. (1997) Chapter 10: An extension of the spatial correlogram and the x -intercept for genetic data. In *Statistical Models for Fluctuating Populations: Patterns and Processes in Time and Space*, O.N. Bjørnstad, Dr Philos. Dissertation, University of Oslo, Oslo.
- Bjørnstad, O.N., Ims, R.A., and Lambin, X. (1999a) Spatial population dynamics: Analysing patterns and processes of population synchrony. *Trends in Ecology and Evolution*, **11**, 427–31.
- Bjørnstad, O.N., Stenseth, N.C., and Saitoh, T. (1999b) Synchrony and scaling in dynamics of voles and mice in northern Japan. *Ecology*, **80**, 622–37.
- Bjørnstad, O. and Bolker, B. (2000) Canonical functions for dispersal-induced spatial covariance and synchrony. *Proceedings of Royal Society London, B.*, **267**, 1787–94.

- Box, G.E.P. and Jenkins, G.M. (1970) *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco.
- Cressie, N. (1993) *Statistics for Spatial Data*, Wiley, New York.
- Deutsch, C.V. and Journel, A.G. (1992) *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, New York.
- Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman and Hall, London.
- Epperson, B.K. (1993a) Recent advances in correlation studies of spatial patterns of genetic variation. *Evolutionary Biology*, **27**, 95–155.
- Epperson, B.K. (1993b) Spatial and space-time correlations in systems of subpopulations with genetic drift and migration. *Genetics*, **133**, 71–27.
- Epperson, B.K. (1995) Fine-scale spatial structure: correlations for individual genotypes differ from those for local gene frequencies. *Evolution*, **49**, 1022–26.
- Epperson, B.K. and Li, T. (1997) Gene dispersal and spatial genetic structure. *Evolution*, **51**, 672–81.
- Green, P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London.
- Hall, P. (1993) On edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation. *Journal of Royal Statistical Society B*, **55**, 291–304.
- Hall, P., Fisher, N.I., and Hoffmann, B. (1994) On the nonparametric estimation of covariance functions. *Annals of Statistics*, **22**, 2115–34.
- Härdle, W. (1990) *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.
- Hilgers, J.W., Reynolds, W.R., Strang, D.E., and McManamey, J.R. (1996) Correlation length used as a predictor of fundamental scale lengths for image characterization. *Optical Engineering*, **35**, 786–93.
- Hyndman, R.J. and Wand, M.P. (1997) Nonparametric autocovariance function estimation. *Australian Journal of Statistics*, **39**, 313–24.
- Jones, M.C., Davies, S.J., and Park, B.U. (1994) Versions of kernel-type regression estimators. *Journal of the American Statistical Association*, **89**, 825–32.
- Journel, A.G. and Huijbregts, C.J. (1978) *Mining Geostatistics*, Academic Press, London.
- Koenig, W.D. (1999) Spatial autocorrelation of ecological phenomena. *Trends in Ecology and Evolution*, **14**, 22–6.
- Lande, R. (1991) Isolation by distance in a quantitative trait. *Genetics*, **128**, 443–53.
- Legendre, P. and Fortin, M.-J. (1989) Spatial pattern and ecological analysis. *Vegetatio*, **80**, 107–38.
- Linhart, Y.B. and Grant, M.C. (1996) Evolutionary significance of local genetic differentiation in plants. *Annual Review of Ecology and Systematics*, **27**, 237–77.
- MacDonald, I.L. and Zucchini, W. (1997) *Hidden Markov and Other Models for Discrete-Valued Time Series*, Chapman & Hall, London.
- Manly, B.F.J. (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, (2nd ed.), Chapman and Hall, London.
- McGraw, J.B. (1995) Patterns and causes of genetic diversity in arctic plants. In *Arctic and Alpine Biodiversity: Patterns, Causes and Ecosystem Consequences*, F.S. Chapin and C. Körner (eds), Springer-Verlag, Berlin. pp. 33–43.
- Myers, R.A., Mertz, G., and Barrowman, N.J. (1995) Spatial scales of variability in cod recruitment in the North Atlantic. *Canadian Journal of Fisheries and Aquatic Science*, **52**, 1849–62.
- Nordal, I. and Iversen, A.P. (1993) Mictic and monomorphic versus parthenogenetic and polymorphic—a comparison of two Scandinavia mountain grasses. *Opera Botanica*, **21**, 19–27.
- Nychka, D. (1995) Splines as local smoothers. *Annals of Statistics*, **23**, 1175–97.
- Oden, N.L. (1984) Assessing the significance of a spatial correlogram. *Geographical Analysis*, **16**, 1–16.

- Oden, N.L. and Sokal, R.R. (1986) Directional autocorrelation: an extension of spatial correlograms to two dimensions. *Systematic Zoology*, **35**, 608–17.
- Ripley, B.D. (1987). *Stochastic Simulation*, Wiley.
- Scott, D.W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York.
- Sokal, R.R. and Jacquez, G.M. (1991) Testing inferences about microevolutionary processes by means of spatial autocorrelation techniques. *Evolution*, **45**, 152–68.
- Sokal, R.R. and Oden, N.L. (1978a) Spatial autocorrelation in biology. I. Methodology. *Biological Journal of the Linnean Society*, **10**, 199–228.
- Sokal, R.R. and Oden, N.L. (1978b) Spatial autocorrelation in biology. II. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society*, **10**, 229–49.
- Sokal, R.R. and Wartenberg, D.E. (1983) A test for spatial autocorrelation analysis using an isolation-by-distance model. *Genetics*, **105**, 219–37.
- Tsay, R.S. (1992) Model checking via parametric bootstraps in time series analysis. *Applied Statistics*, **41**, 1–15.
- Young, G.A. (1994) Bootstrap: more than a stab in the dark? *Statistical Science*, **9**, 382–415.
- Zimmerman, D.L. (1989) Computationally exploitable structure of covariance matrices and generalized covariance matrices in spatial models. *Journal of Statistics and Computer Simulation*, **32**, 1–15.

Biographical sketch

Ottar N. Bjørnstad is a quantitative ecologist at the National Center for Ecological Analysis and synthesis in Santa Barbara. His main research interest in statistical ecology centers on bridging the gap between theory and data, particularly with respect to estimating parameters in ecologically realistic models for abundance data.