Spatial Statistics in Environmental Science

Richard L. Smith [1]

1 Introduction

Spatial statistics is the natural generalization of signal processing to higher dimensions. In traditional signal processing, one has a signal $X(t)$ dependent on a scalar variable t, which may belong to a discrete set or which may be continuous (e.g. the whole real line). Spatial statistics is concerned with cases in which t is a multidimensional index of dimension $d > 1$. In most practical examples, $d = 2$, though much of the basic theory and methodology is the same whatever the dimension. Although the models and methods of spatial statistics have not developed as rapidly as those for one-dimensional signal processing, there have nevertheless been substantial new developments in recent years. Standard references on spatial statistics include the books of Ripley (1981, 1988) and Cressie (1993).

Applications of spatial statistics cover many areas. Much of the original impetus for the subject was driven by geostatistics, e.g. given measurements of the concentration of a mineral at a finite set of positions within a geological field, the aim was to determine the overall volume of the mineral over the whole field. It was in this context that the technique of "kriging" — optimal least squares interpolation over a random spatial field — was originally developed. In recent years, the applications of spatial statistics have increased enormously, with particularly rich applications in the environmental and ecological sciences. A typical problem is the sampling of a pollution field, such as ozone in the atmosphere or toxic chemicals in rivers and lakes. Another example is the use of meteorological measurements in studies of global climate change. In these fields, as in geostatistics, the objective may be to interpolate spatially between measurements, but there are also other objectives which may be quite different. For example, in the context of global climate change, a natural question is to what extent the data support hypotheses of increasing temperature or rainfall, and how the resulting trends, if they exist, vary over the earth's surface. This kind of problem is discussed in section 7. Spatial statistics has also found applications in such diverse fields as sociology — for example, social networks theory — and financial economics: an example of the latter is the term structure of interest rates, where there are two distinct "time" parameters, one the time at which a loan is taken out, the other the

[1]Department of Statistics, University of North Carolina, Chapel Hill, N.C. 27599-3260, U.S.A. Email: rls@email.unc.edu. This work was supported in part by N.S.F. grant DMS-9705166, and by E.P.S.R.C. Visiting Fellowship GR-K99015 at the Isaac Newton Institute.

term of the loan. However, in the present chapter we shall concentrate on environmental applications, which is where the most rapid growth has occurred in recent years.

Sections 2–6 are a review of various concepts and methods in spatial statistics. Section 7 discusses recent developments in spatial trend estimation, motivated by the problem of characterizing the spatial pattern of climate change, but also having applications in a number of other fields.

2 Geostatistics and Kriging

A spatial process will be represented by $Z(s)$, where s varies over a domain $\mathcal{D}$ contained in d-dimensional Euclidean space $\mathbb{R}^d$ for some $d > 1$. Typically, but not necessarily, $d = 2$. As an example, $Z(s)$ might be the concentration of atmospheric ozone taken at a specific place s on the earth's surface (where, for the purpose of the present discussion, we regard the earth's surface as two-dimensional). Suppose also we have a finite number of observations, $z_i = Z(s_i)$ for $i = 1, ..., n$, $s_1, ..., s_n$ denoting the positions of the monitoring sites. Typically in environmental applications (and a contrast with traditional time series analysis), there is no fixed lattice of measuring sites and we regard the variable s as varying over a continuous set $\mathcal{D}$.

The classical "kriging" problem is as follows: given z_i, $i = 1, ..., n$, predict $z_0 = Z(s_0)$ for some new location s_0 which is not one of the given sites $s_1, ..., s_n$. Once this problem is solved, it is easily extended to other problems such as jointly estimating the values of $Z(s)$ at several unmonitored sites s, or estimating a quantity such as $\int_A Z(s)ds$ for some set $A \subseteq \mathcal{D}$.

Models for z_i are traditionally of two types:

$$z_i = \mu + \eta_i, \tag{1}$$

$$z_i = x_i^T \beta + \eta_i, \tag{2}$$

where in each case $\{\eta_i\}$ represent some spatially correlated zero-mean "noise" process. In (1), μ is a single (usually unknown) parameter assumed to be constant at all points on the surface, while in (2), the mean at a specific point is assumed to depend on a given set of covariates x_i through a linear regression model with unknown parameters β. Traditionally (1) is described as the "ordinary kriging" problem and (2) as "universal kriging".

Covariance assumptions on $\eta(\cdot)$ may be represented through

$$\mathrm{Cov}\{\eta(s), \eta(s')\} = C(s, s'), \tag{3}$$

for some covariance function $C(\cdot, \cdot)$.

Various homogeneity assumptions may be made on the covariance function. For example, if C is of the form

$$C(s, s') = C_0(s - s') \tag{4}$$

for some C_0, then the process is described as *stationary*. It captures the property (the obvious generalization of stationarity in time series analysis) that the dependence between two sites s and s' depends only on the (vector) distance between the sites, $s - s'$.

Instead of the covariance function, it is common in spatial statistics to work instead with the *semivariogram function* $\gamma(\cdot)$, defined by

$$\mathrm{Var}\{\eta(s) - \eta(s')\} = 2\gamma(s - s'). \tag{5}$$

For somewhat odd historical reasons, the left hand side of (5) is called the *variogram*, and the function γ the semivariogram. One motivation for considering the (semi)variogram rather than the covariance function is that the assumption (5), also known as *intrinsic stationarity*, is actually somewhat weaker than (4), in that there are models for which the variogram exists when the covariance function does not. For the applications considered in this chapter, however, it will be good enough to assume that the covariance function exists, and then it does not really matter whether the covariance function or the variogram is used to characterize the process.

Denoting the argument of C_0 or γ as h, if either $C_0(h)$ or $\gamma(h)$ depends only on $||h||$, i.e. the length of h, usually measured via the usual Euclidean metric, the process is said to be *isotropic*. Sometimes the word *homogeneous* is used to describe a process which is both stationary and isotropic.

2.1 Kriging

Rewrite the model (2) in the form

$$\mathbf{Z} = X\beta + \eta, \tag{6}$$

where $\mathbf{Z}^T = (\, z_1 \quad z_2 \quad \ldots \quad z_n \,)$, X is the $n \times p$ matrix of covariates, β is a p-dimensional vector of unknown regression coefficients, and η is a vector of random errors with mean 0 and covariance matrix of the form $C = \alpha V$, where $\alpha > 0$ is allowed to be an unknown positive scalar but the matrix V is assumed known. In most cases, we also assume that η has a multivariate normal distribution. As stated, with the regression term $X\beta$, this is the universal kriging problem, but ordinary kriging is a special case in which $X\beta$ is replaced by $\mathbf{1}\mu$, $\mathbf{1}$ being an n-dimensional vector of ones and μ a fixed unknown constant. Suppose we wish to predict a value

$$z_0 = x_0^T\beta + \eta_0, \tag{7}$$

in which the covariates x_0 at a new site are given, and η_0 is a random variable with mean 0, variance αv_0 and covariance with η given by $\mathrm{E}\{\eta^T\eta_0\} = \alpha w$ with the scalar v_0 and the vector w both known. Consider a linear predictor of the form $\hat{z}_0 = \lambda^T\mathbf{Z}$ where λ satisfies the constraint

$$X^T\lambda = x_0. \tag{8}$$

The rationale for the constraint (8) is that it justifies the reduction

$$\hat{z}_0 - z_0 = \lambda^T(X\beta + \eta) - x_0^T\beta - \eta_0 = \lambda^T\eta - \eta_0,$$

or in other words, the prediction error does not depend on the unknown quantity β. There are now several ways to go about the solution:

1. Find λ to minimize $\mathrm{E}\{(\hat{z}_0 - z_0)^2\}$ subject to the constraint (8). This does not involve any normality assumption, since the formulation of the problem depends solely on first- and second-order moments. The most direct solution is via the method of Lagrange multipliers, leading to

$$\lambda = \{(x_0 - X^TV^{-1}w)^T(X^TV^{-1}X)^{-1}X^T + w^T\}V^{-1}, \tag{9}$$

with an accompanying (complicated) expression for the variance of $\hat{z}_0$ given $\mathbf{Z}$. See, for example, Ripley (1981) or Cressie (1993).

2. A Bayesian solution (assuming normality) is to fix the improper prior distribution $\pi(\beta) \equiv 1$, calculate the posterior density $\pi(\beta|\mathbf{Z})$, and then

$$\hat{z}_0 = \int \mathrm{E}\{z_0|\mathbf{Z}, \beta)\}\pi(\beta|\mathbf{Z})d\beta.$$

3. Suppose $\mathbf{Z}_1 = A\mathbf{Z}$ is an $(n-p)$-dimensional vector of linearly independent contrasts, i.e. linear functions of $\mathbf{Z}$ whose distribution does not depend on β. General vector space theory implies that such a matrix A must exist, though it may not be so easy to calculate explicitly. Also let $\hat{\beta} = (X^TV^{-1}X)^{-1}X^TV^{-1}\mathbf{Z}$ denote the generalized least squares estimator of β. It is easily checked that the distribution of $z_0 - x_0^T\hat{\beta}$ does not depend on the true value of β. Then $\hat{z}_0 - x_0^T\hat{\beta}$ is the conditional expectation of $z_0 - x_0^T\hat{\beta}$ given $\mathbf{Z}_1$.

The equivalence of formulations 2 and 3 is an interesting connection between Bayesian and conditional inference which has been noted in other contexts — see the discussion of REML estimation in the next subsection.

These approaches to kriging all make full allowance for the fact that the regression coefficient β are *a priori* unknown, but they make no allowance at all for the fact that V is also typically unknown, which arguably, is the more important problem! So let us now turn to consideration of this feature.

2.2 Estimation of V

The most common estimation approaches assume that the process is stationary and isotropic. For methods getting away from this assumption, see section 3.

The first step is very often a plot to determine the shape of either the covariance or the semivariogram function. Assuming the process $Z(s)$ has common mean, a common estimator for the semivariogram γ based on a finite number of observations $s = s_1, ..., s_n$ is of the form

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|}\sum_{(s_i,s_j)\in N(h)}\{Z(s_i) - Z(s_j)\}^2, \tag{10}$$

where $N(h)$ denotes a collection of (s_i, s_j) pairs whose Euclidean distance lies within a given neighborhood of h, and $|\cdot|$ denotes cardinality.

The estimate (10) is sometimes criticized for being overly sensitive to outliers and there are various "robust" alternatives, for example, the Cressie-Hawkins estimator (Cressie, 1993)

$$2\tilde{\gamma}(h) = \frac{1}{0.457 + \frac{0.494}{n}} \left\{ \frac{1}{|N(h)|} \sum_{N(h)} |Z(s_i) - Z(s_j)|^{1/2} \right\}^4, \tag{11}$$

which is also an approximately unbiased estimator when the data are normally distributed, but which is less affected by outliers than (10).

The second step in the estimation procedure is to choose from a family of positive-definite covariance functions (or, equivalently, negative-definite semivariogram functions). Simple estimators such as (10) and (11), when viewed as a function of h, do not have the negative-definiteness property which is a necessary condition for a legitimate semivariogram function. Therefore, in most cases, we choose from a parametric family which does have this property.

Formulating the problem in terms of covariances rather than semivariograms, a general condition for a positive-definite covariance in a stationary process is that it be expressible in the form

$$C_0(h) = \int \cos(\omega^T h) G(d\omega), \tag{12}$$

where G is a non-negative measure on $\mathbb{R}^d$. In isotropic cases, where the vector argument h is replaced by a scalar argument t, the formula reduces to

$$C_0(t) = \int_{(0,\infty)} Y_d(\omega t) \Phi(d\omega), \quad 0 < t < \infty, \tag{13}$$

with Φ a non-negative measure on $(0, \infty)$. In (13), Y_d is given (depending on dimension d) by

$$Y_d(t) = \left(\frac{1}{t}\right)^{(d-2)/2} \Gamma\left(\frac{d}{2}\right) J_{(d-2)/2}(t),$$

$\Gamma(\cdot)$ being the usual gamma function and $J_{(d-2)/2}(\cdot)$ a Bessel function of the first kind of order $(d-2)/2$. See, for example, Ripley (1981) or Cressie (1993).

In practice, there are a number of standard families of covariance functions (or, equivalently, semivariograms) which are consistent with (13). Examples include:

Spherical model:

$$\gamma_0(t) = \begin{cases} 0, & t = 0, \\ c_0 + c_1 \left\{ \frac{3}{2}\frac{t}{R} - \frac{1}{2}\left(\frac{t}{R}\right)^3 \right\}, & 0 < t \le R, \\ c_0 + c_1, & t \ge R, \end{cases} \tag{14}$$

Exponential model:

$$\gamma_0(t) = \begin{cases} 0, & t = 0, \\ c_0 + c_1 \left(1 - e^{-t/R}\right), & t > 0, \end{cases} \tag{15}$$

Gaussian model:

$$\gamma_0(t) = \begin{cases} 0, & t = 0, \\ c_0 + c_1 \left(1 - e^{-(t/R)^2}\right), & t > 0, \end{cases} \tag{16}$$

Matérn model: Best defined in terms of the covariance function C_0 as

$$C_0(t) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\sqrt{\nu}t}{R}\right)^{\nu} \mathcal{K}_\nu \left(\frac{2\sqrt{\nu}t}{R}\right), \tag{17}$$

where $\nu > 0$ is a shape parameter and $\mathcal{K}_\nu$ is the modified Bessel function of the third kind of order ν (Abramowitz and Stegun 1964, Chapter 9). The special cases $\nu = \frac{1}{2}$ and $\nu \to \infty$ correspond respectively to the exponential and Gaussian models, (15) and (16).

In each of (14)–(17), $R > 0$ is a scale parameter is known as the *range*, and in (14)–(16), if $c_0 \neq 0$ it is known as the *nugget*. This reflects the commonly observed property that, even at very small distances, observed variograms are non-negligible, which is often interpreted to reflect measurement errors in the observations rather than real discontinuities in the surface being measured.

Discrimination among parametric models is often carried out by visual assessment based on the sample semivariogram $\hat{\gamma}(\cdot)$ or $\tilde{\gamma}(\cdot)$, but may also be carried out (after fitting the model) by more formal criteria such as likelihood ratio tests, AIC, BIC, etc.

The third step in the estimation procedure is the estimation of parameters of the assumed covariance (or variogram) model. One relatively simple, and reasonably efficient, technique for this is Cressie's weighted least squares procedure (Cressie, 1993): given a sample variogram $\hat{\gamma}(h)$ evaluated at a finite number of values of h, say $h_1, h_2, ...$, and a model $\gamma(h;\theta)$ depending on unknown parameters θ, choose θ to minimize

$$\sum_j |N(h_j)| \left\{\frac{\hat{\gamma}(h_j)}{\gamma(h_j;\theta)} - 1\right\}^2. \tag{18}$$

The method is not dependent on a particular sample estimator; for example, $\hat{\gamma}(\cdot)$ from (10) may be replaced by $\tilde{\gamma}(\cdot)$ from (11). This method has the advantage of being relatively straightforward to calculate, requiring a nonlinear optimization but no complicated likelihood evaluation. A practical disadvantage is that there is no easy way to obtain standard errors for the estimators, or tests of hypotheses about the parameters.

Most of the alternatives to this method are based on some form of likelihood procedure, assuming a Gaussian process:

• *Maximum likelihood estimation* (Kitanidis 1983, Mardia and Marshall 1984). This is more complicated than Cressie's method because the evaluation of the exact likelihood is appreciably harder than (18), but it is computationally feasible for reasonably sized problems. If there are n data points, likelihood evaluation requires storage and inversion of a $n \times n$ covariance matrix; the author has successfully applied this for n up to 500, but there would clearly be problems if n were of the order of several thousand.

• *Restricted maximum likelihood (REML) estimation* (Cressie, 1993). This is an alternative to maximum likelihood, especially well adapted to models of the form of (6), for which it uses a likelihood function for θ based on a set of contrasts of $\mathbf{Z}$, orthogonal to the design matrix X, whose distribution is unaffected by β. Although asymptotically equivalent to maximum likelihood, the method is generally believed to have superior properties in small samples, especially when the dimension of β is large. As shown originally by Harville (1974), the method is also equivalent to a simple form of Bayesian analysis in which the parameter β is given a uniform prior density, though the treatment of the θ parameter is not Bayesian.

• *Bayesian methods* (Le and Zidek 1992, Handcock and Stein 1993, Brown *et al.* 1994), in which all the unknown parameters are given prior distributions and a joint posterior distribution is calculated, have become much more popular in recent years, though they are more complicated computationally than ML or REML estimation.

An example of these estimation methods if shown in Fig. 1. This is based on the Texas aquifer data set of Cressie (1993), in which the underground water level was measured at various places in Texas. Cressie's original analysis used an intrinsically stationary model without directly adjusting for any deterministic spatial trend. In the present analysis, a linear trend was incorporated through the regression function in (2), and the residuals modeled as a stationary, isotropic process. The standard (10) and robust (11) variograms of the residuals are shown, along with an exponential semivariogram model fitted by each of the weighted least squares, maximum likelihood and REML methods. The results appear to show good consistency among the different methods.

The fourth and final step in the estimation procedure is the application to kriging. In traditional geostatistical applications, once the covariance matrix V was estimated by any of the methods we have described, this was simply treated as a known matrix for the calculation of kriging formulae and their prediction variances. However, as already noted, it is somewhat illogical to adapt the kriging procedure for the fact that β is unknown without making a similar allowance for the unknown parameters in V. Two possible approaches are:

• *Corrections based on the delta method.* These usually use the standard kriging formulae for point predictions but correct the prediction variances,

using Taylor expansions to represent the effect of parameter uncertainty. See Zimmerman and Cressie (1992), Harville and Jeske (1992).

• *Fully Bayesian approach.* It is straightforward to formulate the problem from a Bayesian viewpoint, since it treats (β, θ, z_0) (where β are the regression parameters, θ the unknown parameters of the covariance matrix, and $z_0 = Z(s_0)$ the unknown quantity being predicted) as a random vector, and calculates the conditional distribution of z_0 given observed data Z after integrating out the effect of β and θ. As already noted, in the special case that θ is known and β has a uniform (improper) prior distribution, this is equivalent to the standard kriging formulae.

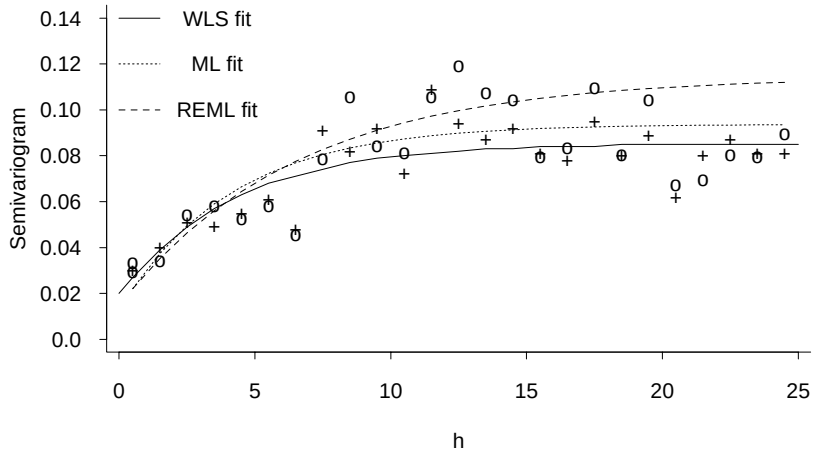

Fig. 1. Standard (+) and robust (o) estimates of the sample semivariogram with fitted exponential semivariogram function by the weighted least squares, maximum likelihood and restricted maximum likelihood methods.

3 Nonstationary models

The approach described in section 1 is based on an assumption that the spatial random field is stationary and isotropic. In the original geophysical applications which motivated the development of the field, this assumption was often justified by the fact that with sparse data, there was no reasonable alternative. A further point is that many geostatistical applications involved only one measurement at each site (or equivalently, only one replication of the random field) so there was no way of determining the complete spatial covariance function without some kind of stationarity assumption. In modern environmental applications, however, there are very often enough monitoring stations to go beyond such assumptions, and with multiple observations per site, it is also possible to estimate the covariance between any pair of sites without assuming

stationarity across the field. Another consideration is that very often, simple topography makes a stationarity assumption implausible. Therefore, there are by now many reasons to go beyond a stationary model.

In spite of this obvious need for nonstationary models, however, there is not as yet a wide variety of approaches to the problem. In the present section we concentrate on one particular approach, pioneered by Sampson and Guttorp (1992) and also developed by, among others, Mardia and Goodall (1993), Guttorp *et al.* (1994) and Smith (1996).

The idea is a "deformation approach" to nonstationarity: we assume the observed process is nonstationary, but that it can be deformed into a stationary (and, in most applications, isotropic) process by some nonlinear map. For a spatial process $Z(s)$ with constant mean, defined for sites s, t within some domain $\mathcal{D}$, define the *dispersion function*

$$D(s,t) = \mathrm{E}\left[\{Z(s) - Z(t)\}^2\right]$$

for each pair of sites (s,t). We look for models of the form

$$D(s,t) = 2\gamma_0(||f(s) - f(t)||) \tag{19}$$

where f is some nonlinear function on $\mathcal{D}$ and γ_0 is the semivariogram of a stationary, isotropic process. In the terminology developed by Sampson and Guttorp, the original "geographic space", or G-space, in which the observations are located, is transformed by the function f into a "dispersion space", or D-space, and in that space the process is stationary and isotropic.

To estimate such a model, it is not sufficient to have only one observation per location, because without any stationarity assumption, it is not possible to measure the covariance between any pair of points. However, virtually all the practical applications of this methodology have been in contexts for which there is no shortage of available data to estimate the covariances required.

Guttorp and Sampson, and their co-authors, have developed a variety of ingenious but somewhat *ad hoc* fitting techniques for these models. For example, Guttorp *et al.* (1994) proposed an estimation scheme in which f and γ_0 were chosen to minimize

$$\sum_i \sum_j \left(\frac{d_{ij} - D_{ij}}{d_{ij}}\right)^2 + \lambda\{J(f^{(1)}) + J(f^{(2)})\}, \tag{20}$$

where $f^{(1)}$ and $f^{(2)}$ are the two coordinate functions of $f: \mathbb{R}^2 \to \mathbb{R}^2$, and

$$J(f^{(j)}) = \int\int \left\{ \left(\frac{\partial^2 f^{(j)}}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 f^{(j)}}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 f^{(j)}}{\partial y^2}\right)^2 \right\} dxdy. \tag{21}$$

Note that we are representing a generic point of $\mathbb{R}^2$ by (x,y) rather than s as previously. In (20) , D_{ij} is the observed dispersion between sites i and

j as determined empirically from the data, d_{ij} is the model-based dispersion determined by the functions f and γ_0, $\lambda > 0$ is a smoothing parameter, and $J(f^{(j)})$, $j = 1, 2$, is a "bending energy" functional whose presence ensures that the function f is not allowed to become too irregular. This kind of functional is often used as a penalty function in spline-based approaches to smoothing and interpolation.

An alternative approach (Mardia and Goodall 1993, Smith 1996) is to choose f and γ_0 to minimize the profile negative log likelihood function

$$L = \frac{N}{2} \log |\Sigma| + \frac{N-1}{2} \operatorname{tr}\left(\Sigma^{-1}\hat{\Sigma}\right), \tag{22}$$

where we assume there are N replications of the field, $\hat{\Sigma}$ is the sample covariance matrix and Σ is the model-based covariance matrix.

To make this approach work effectively, we would really like to represent both γ_0 and f as parametric functions. Parametric models for γ_0 have already been discussed, but an alternative approach, motivated by the representation (13), is to represent the covariance function C_0 in the form

$$C_0(h) = \sum_{c=1}^{C} \phi_c J_0(\omega_c h) \tag{23}$$

for a fixed C and positive constants $\phi_1, ..., \phi_C, \omega_1, ..., \omega_C$. The idea underlying (23) is that it forms a discrete approximation to the integral model (13). Note that the function Y_d in (13) reduces to J_0 when $d = 2$, as we assume throughout this section.

The other issue involved in parametrizing the model is to obtain a finite-parameter representation for f. Assuming again that we are working in dimension $d = 2$, f can be represented as $(f^{(1)}, f^{(2)})$, where $f^{(1)}$ and $f^{(2)}$ are scalar functions of location $s = (x, y)$. There are by now many approaches to non-linear function reconstruction that use an expansion in terms of basis functions: for example, thin-plate splines, radial basis functions (RBFs) and wavelets all use representations of this form. In dimension 2, the thin-plate spline and RBF approaches coincide, and lead to $f^{(1)}$ and $f^{(2)}$ being represented as linear combinations of basis functions of the form

$$\eta_i(x, y) = r^2 \log r,$$

where $r = \left\{(x - x_i)^2 + (y - y_i)^2\right\}^{1/2}$, the distance between the current location (x, y) and the i'th "center" (x_i, y_i). The model is completely determined once we specify the number and positions of the centers. We then write

$$f^{(j)}(x, y) = \sum_{i=1}^{I} \delta_i^{(j)} \eta_i(x, y), \quad j = 1, 2,$$

where I is the number of centers and $\{\delta_i^{(j)},\ i = 1, ..., I,\ j = 1, 2\}$ are coefficients to be determined.

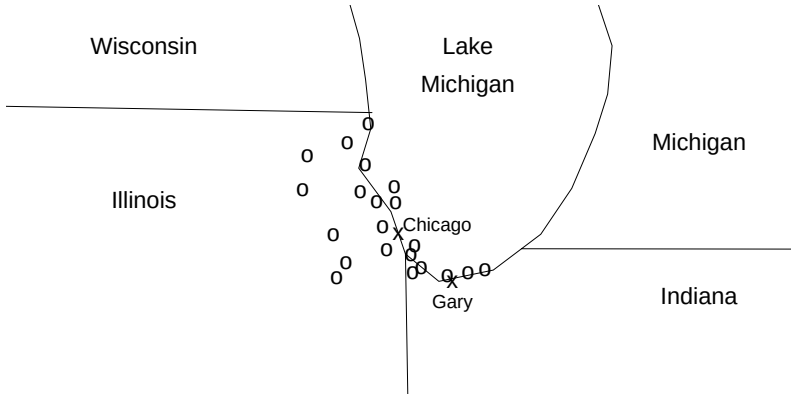

Fig. 2. Map of ozone stations: G-space.

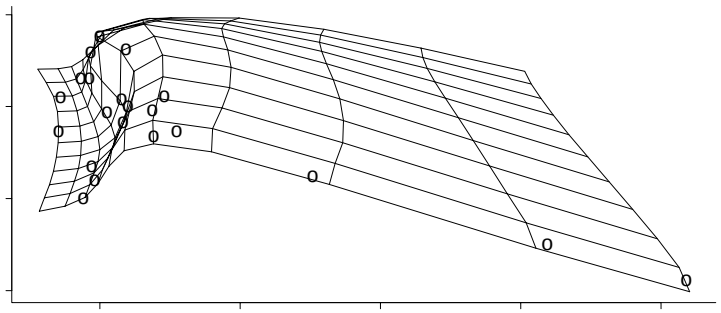

Fig. 3. Map of ozone stations: D-space.

As an example of this approach (taken from Smith, 1996), Fig. 2 shows a map of ozone stations in the neighborhood of Chicago in the original "G-space". The bulk of the stations shown are in the city of Chicago or surburban Illinois, but the three lower right stations are near the city of Gary, Indiana, a heavily industrialized region. Sample correlations among ozone measurements show that these three stations have much lower correlations (with each other, as well as with the remaining stations in the figure) than the rest of the stations, so we would not expect a stationary model to hold. Throughout the

examples in this section, we work with spatial correlations rather than spatial covariances to avoid having to deal with the fact that the variances may also vary from site to site. Fig. 3 indeed shows that (one reconstruction of) the "D-space" in this example is highly distorted, with the three outlying stations drawn out into positions very distant from the remainder.

As an example of the effect of the transformation on the sample dispersion, Fig. 4 shows the dispersions in the original G-space, with dispersions involving one of the three outlying stations distinguished by a separate plotting symbol. It is evident that the plot cannot easily be represented by a smooth curve. The transformed picture (Fig. 5), together with the fitted Matérn variogram function, is much cleaner.

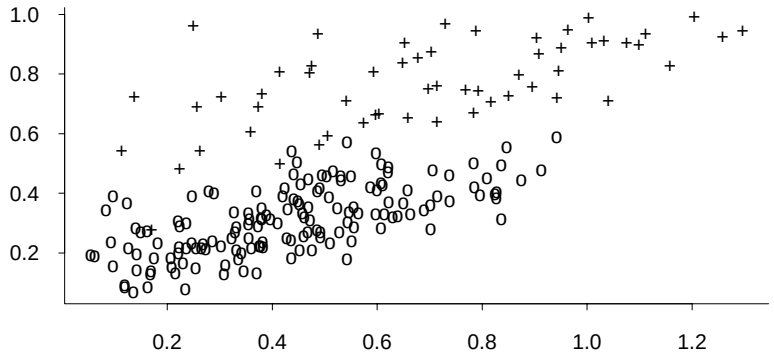

Fig. 4. Dispersion plots in G-space for the ozone example. For each pair of stations, the distance between the stations is plotted along the horizontal axis and the dispersion is plotted along the vertical axis. For any dispersion calculation involving one of the three outlying stations, the plotted point is +; for the remainder, it is o.

A second example is based on temperature averages for stations in a subset of the United States "Historical Climatological Network". The top plot in Fig. 6 shows the locations of the stations superimposed on a map of the country. The bottom plot shows the transformed D-space. In this example, the striking feature of the plot is that stations in the southwestern states (California, Nevada, Arizona) have been pulled far away from the rest of the plot, while stations in the northwest (Washington, Oregon) are drawn closer to the rest of the country. However, the corresponding "before and after" variogram plots (Fig. 7) so not have nearly so clear-cut an interpretation as Fig. 5. In this case, the fitted Matérn curve in D-space still does not come close to fitting the whole dispersion function, and one is led to suspect that the

transformed model still does not fit the data very well. Further investigation based on the covariance function (23) improves the fit (Smith 1996), but still without adequately fitting the whole of the data.

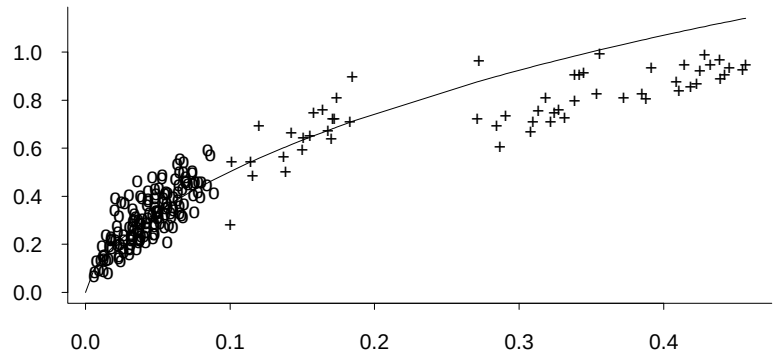

Fig. 5. Figure 4 redrawn in D-space, with fitted Matérn variogram curve.

3.1 Alternative approaches to nonstationary covariances

Although the deformation approach to nonstationary spatial processes is the one which has been most extensively developed, it is far from being the only approach to this problem. One fairly straightforward approach is the "moving window" method due to Haas (1990, 1995). According to this, kriging at a particular location is based only on a subset of the monitoring stations within a given distance of the location for which a prediction is needed. In a space-time context, the window is defined in time as well as space, thus allowing for temporal nonstationarity as well. The actual size of the window is chosen by some form of cross-validation scheme. This approach is relatively easy to apply, in part because it avoids the complications inherent is specifying a full model for the nonstationary case. On the other hand, the lack of a fully specified model is a disadvantage in some contexts.

Another approach is the *orthogonal expansion approach*, discussed by Nychka and Saltzman (1998), Holland *et al.* (1999). This is based on

$$C(s,t) = \sigma(s)\sigma(t)\left[\rho e^{-||s-t||/R} + \sum_{\nu=1}^{m} \lambda_\nu \psi_\nu(s)\psi_\nu(t)\right], \tag{24}$$

where $\sigma(\cdot)$ is a spatially varying standard deviation, $0 \leq \rho \leq 1$, λ_ν is a non-negative weight, and ψ_ν is the νth eigenfunction in some orthogonal function

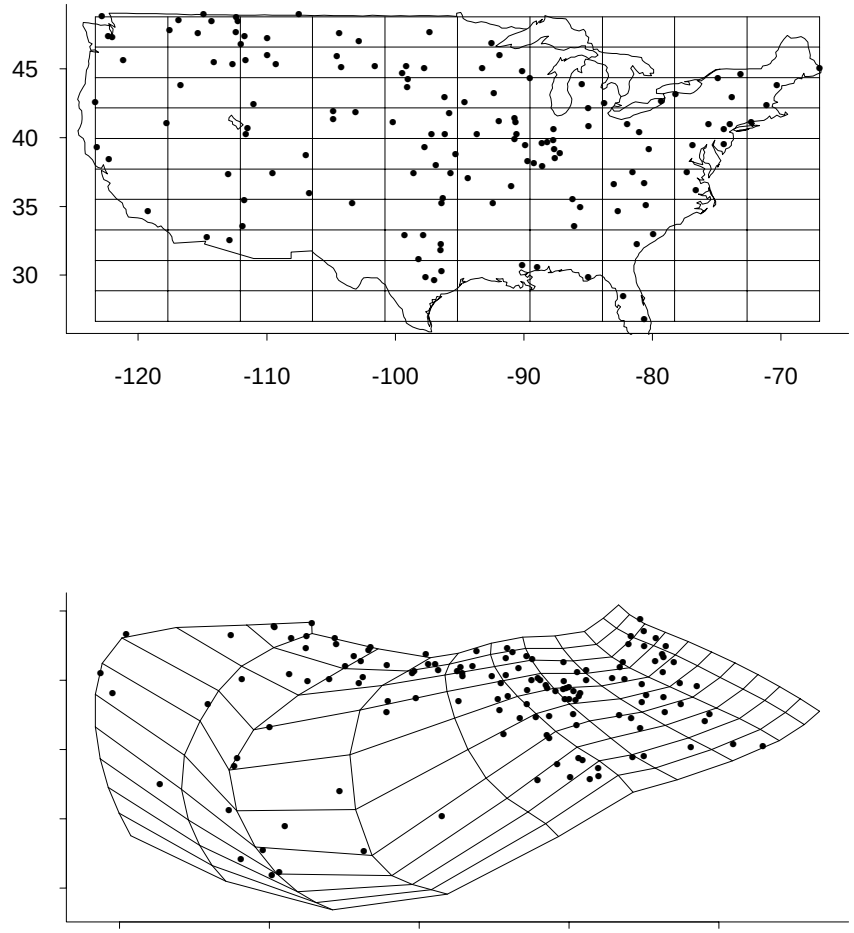

Fig. 6. G-space and D-space for climatological example

expansion. As Nychka and Saltzman point out, the representation (24) is equivalent to a representation of the random field as

$$Z(s) = \sigma(s) \left\{ \rho Z_0(s) + \sum_{\nu=1}^{m} a_\nu \lambda_\nu^{1/2} \psi_\nu(s) \right\}, \tag{25}$$

where Z_0 is a stationary isotropic random field and $a_1, ..., a_m$ are independent

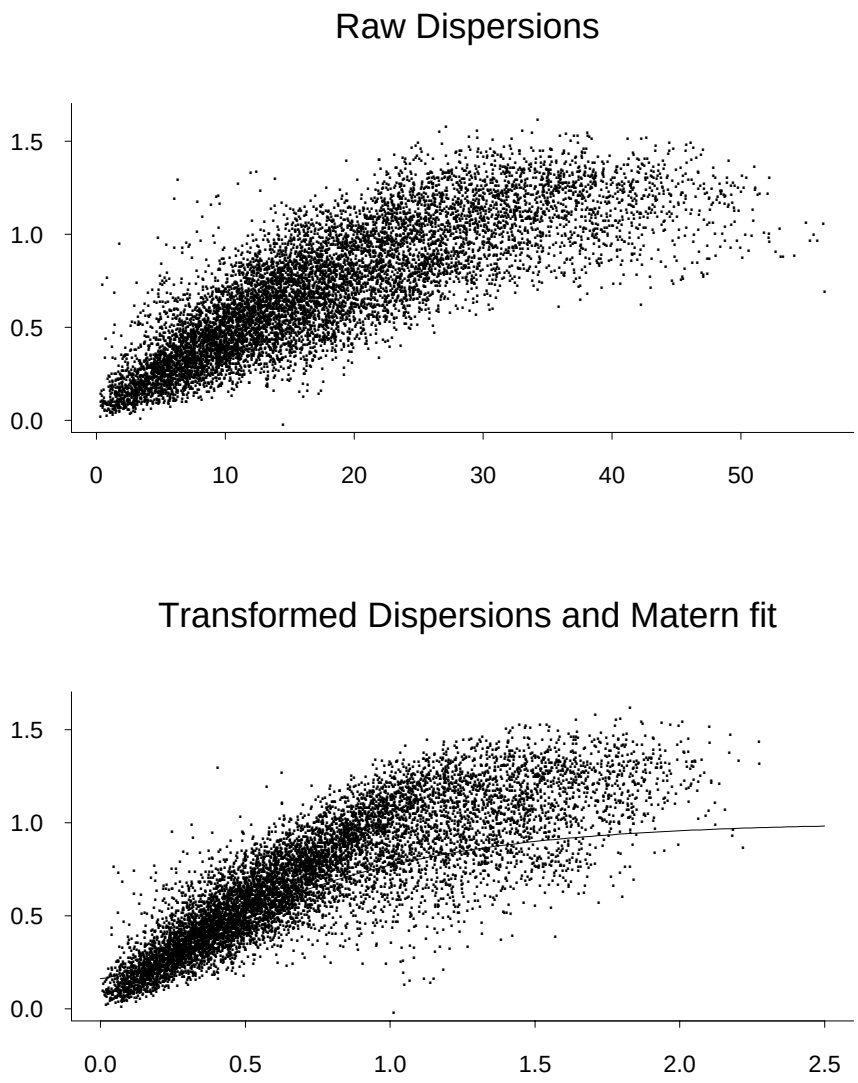

Fig. 7. Top plot: sample dispersion function for climatological data in G-space. Bottom plot: same in D-space, with fitted Matérn curve.

(of each other and of Z_0) standard normal random variables. This is therefore a hybrid between the geostatistical approach based on stationary random fields, and an approach known as empirical orthogonal functions, which is popular in atmospheric sciences. Presumably, the approach is not restricted to the exponential covariance function in (24) but other forms of stationary isotropic

covariance function could also be used. At the present time, the whole approach based on expansions of the form of (24) has not been developed very extensively.

4 Models defined by conditional probabilities

An entirely different approach to modeling spatial fields is through families of conditional distributions for the observation at a site given its neighbors. Such models are most naturally defined for a discrete set of locations, though they are often applied in situations where the underlying random field is defined continuously in space. The whole approach stems from Besag (1974), though there have been many extensions and variations on the approach in recent years. A small sample of recent papers are Besag *et al.* (1995), Waller *et al.* (1997), Best *et al.* (1998), Diggle *et al.* (1998), Wolpert and Ickstadt (1998), Besag and Higdon (1999). In the present discussion, we do not attempt anything like a complete survey of recent developments, but will concentrate on outlining the fundamental ideas, stemming from Besag (1974).

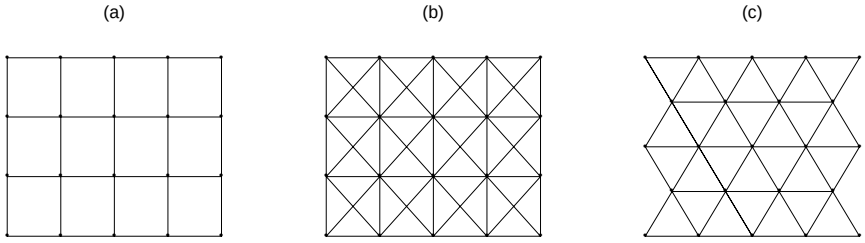

Fig. 8. Examples of regular lattices. The vertices of the graph represent spatial locations, and two vertices are said to be neighbors if there is edge of the graph joining them.

The starting point is that one considers data defined on a discrete lattice with an underlying neighborhood structure. The lattice may have some regular shape as in Fig. 8, though often a lattice is defined simply as a graph with no regular structure. Where the true field is distributed continuously in space, the spatial locations are usually aggregated into discrete cells so that such a model can be applied.

A model is defined by specifying the conditional distribution at a particular site given its neighbors. For example, in the *autologistic model*, each site value X_i is either 0 or 1, and satisfies

$$\begin{aligned} \Pr\{X_i = 1 \mid X_j = x_j,\ j \neq i\} &= \Pr\{X_i = 1 \mid X_j = x_j,\ j \in N_i\} \\ &= \frac{\exp\left(\alpha_i + \sum_{j \in N_i} \beta_{ij} x_j\right)}{1 + \exp\left(\alpha_i + \sum_{j \in N_i} \beta_{ij} x_j\right)} \end{aligned} \tag{26}$$

where N_i is the set of neighbors of the site i, α_i is a coefficient for each site, and β_{ij} is an interaction coefficient between neighboring sites i and j.

A corresponding model for normally distributed systems is the *autonormal* model, defined as follows: the conditional distribution of X_i given $X_j,\ j \neq i$ is normal with variance σ^2 and mean

$$\mu_i + \sum_{j \in N_i} \beta_{ij}(X_j - \mu_j).$$

Thus μ_i is the mean at site i and β_{ij} is again a pairwise interaction component defined for neighboring sites i and j.

It should be noted that the autonormal model just defined is not the same as

$$X_i = \mu_i + \sum_{j \in N_i} \beta_{ij}(X_j - \mu_j) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \text{ (independent)},$$

which is called the *simultaneous equation model* and corresponds to a quite different joint distribution of the $\{X_i\}$. Besag (1974) discussed this distinction in detail.

A key question with models of this form is whether the family of conditional probabilities is consistent with some set of joint probabilities on all the random variables. If it is not, then clearly the model is not well defined. Both the autologistic and autonormal models are examples of a *Markov random field* (MRF), and the question of whether they are well-defined models is answered through a very general result for MRFs known as the Hammersley-Clifford theorem (Besag 1974, Clifford 1990). For pairwise interaction models such as logistic and autonormal, the answer essentially reduces to the statement that both the neighborhood structure and the interaction coefficients must be symmetric: $j \in N_i$ if and only if $i \in N_j$, and $\beta_{ij} = \beta_{ji}$. For example, in the autologistic case, the joint density of $\mathbf{X} = \{X_i\}$, evaluated at $\mathbf{x} = \{x_i\}$, is of the form

$$p(\mathbf{x}) \propto \exp\left(\sum_k \alpha_k x_k + \frac{1}{2}\sum_j \sum_{k \in N_j} \beta_{jk} x_j x_k\right), \tag{27}$$

while in the autonormal case, the corresponding joint density is

$$p(\mathbf{x}) = (2\pi\sigma^2)^{-1/2}|B|^{1/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_j \sum_k (x_j - \mu_j) b_{jk} (x_k - \mu_k)\right\}, \tag{28}$$

where

$$b_{jk} = \begin{cases} 1 & \text{if } j = k, \\ -\beta_{jk} & \text{if } j \in N_i, \\ 0 & \text{otherwise.} \end{cases}$$

It is readily checked that (27) implies (26) , and that (28) implies the autonormal model. Note that (28) is equivalent to the statement that the $\{X_i\}$ are multivariate normally distributed with covariance matrix $\Sigma = B^{-1}$.

4.1 Estimation of MRF models

(i) *Exact MLE*

Exact maximum likelihood is usually feasible only for Gaussian processes. To see why, note that $p(\mathbf{x})$ in (27) is defined only up to an unknown constant of proportionality, and direct maximum likelihood is not possible without evaluating that constant. However, exact calculation of the constant can only be performed by summing (27) over all possible states of the system, which is impossible unless the system is extremely small. On the other hand, in (28) , we are able to evaluate the constant analytically, so in this case it is possible to calculate the likelihood function exactly. Virtually all non-Gaussian cases are like (27) , in that the model is specified up to an unknown constant of proportionality, but there is no direct method of evaluating the constant.

(ii) *Maximum pseudo-likelihood*

Besag (1975) proposed estimating the unknown parameters of the model, θ say, by maximizing the quantity

$$PL(\theta) = \prod_i p(X_i \mid X_j,\ j \in N_i;\ \theta), \tag{29}$$

which he called the pseudo-likelihood. This has the advantage of being easy to calculate, and behaving in many respects like a likelihood function, though it is not equivalent to the likelihood function, even asymptotically, and in some contexts the maximum pseudo-likelihood estimates are much less efficient than the maximum likelihood estimates. In Gaussian cases, it is possible to compare the two methods directly (Besag and Moran 1975, Besag 1977). The method has fallen under something of a cloud in recent years, partly because of the growing popularity of simulation-based estimation methods such as the Gibbs sampler, though it remains of interest as a theoretical technique (see e.g. Comets, 1992).

(iii) *Simulated maximum likelihood estimators*

The idea behind this was proposed by Penttinen (1984), and extended by Geyer and Thompson (1992). Suppose we have a model of the form

$$p(\mathbf{x};\theta) = C(\theta)F(\mathbf{x};\theta),$$

where $F(\mathbf{x};\theta)$ is a specified function of data $\mathbf{x}$ and unknown parameter θ, and $C(\theta)$ is a normalizing constant, in principle computable by summing or integrating over all possible values of $\mathbf{x}$, but in practice not computable. Our objective is to calculate a simulation-based approximation to the maximum likelihood estimate for a particular realization of the Markov random field, $\mathbf{X}$ say.

Suppose we generate a Monte Carlo sample $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(M)}$ of realizations of the random field $\mathbf{X}$ for some particular value of θ, say θ_0. It is not essential

that $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(M)}$ be independent. Then for $1 \leq m \leq M$,

$$\begin{aligned} \mathrm{E}_{\theta_0}\left\{\frac{F(\mathbf{X}^{(m)};\theta)}{F(\mathbf{X}^{(m)};\theta_0)}\right\} &= \sum_{\mathbf{x}} \frac{F(\mathbf{x};\theta)}{F(\mathbf{x};\theta_0)} \cdot C(\theta_0)F(\mathbf{x};\theta_0) \\ &= C(\theta_0)\sum_{\mathbf{x}} F(\mathbf{x};\theta) \\ &= \frac{C(\theta_0)}{C(\theta)}. \end{aligned} \tag{30}$$

If the distribution of $\mathbf{x}$ is continuous, then the sum in (30) is replaced by an integral. Based on (30) , therefore,

$$\frac{1}{M}\sum_{m=1}^{M} \frac{F(\mathbf{X}^{(m)};\theta)}{F(\mathbf{X}^{(m)};\theta_0)} \cdot \frac{F(\mathbf{X};\theta_0)}{F(\mathbf{X};\theta)} \tag{31}$$

is an unbiased estimate of the likelihood ratio of θ_0 to θ, in other words

$$\frac{C(\theta_0)F(\mathbf{X};\theta_0)}{C(\theta)F(\mathbf{X};\theta)}.$$

An estimator of θ defined so as to minimize (31) , then, may be regarded as a simulated maximum likelihood estimator, and provided M is sufficiently large, may be expected to be a good approximation to the true MLE.

Note that as the method is commonly applied, the sample $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(M)}$ is generated only once, with a fixed θ_0, and then (31) is treated as a deterministic function of θ. However, partly as a check on the simulation-sensitivity of the method, and also in some cases to speed up the convergence to the MLE, it is possible to update θ_0 during the procedure, for example, minimizing (31) to get an initial estimate $\hat{\theta}^{(1)}$, then taking $\theta_0 = \hat{\theta}^{(1)}$ and repeating the process to get an estimate $\hat{\theta}^{(2)}$, and so on.

Many practical issues are raised by this method — for example, the choice of M, and the method of simulation used to generate the individual $\mathbf{X}^{(m)}$ fields. Many modern ideas such as the Gibbs sampler (Geman and Geman 1984), auxiliary variable methods (Swendsen and Wang 1987), Metropolis-Hastings algorithms (Besag *et al.* (1995) reviewed all of these methods) and perfect simulation (Propp and Wilson, 1996) have been developed in recent years for simulation from random fields, but we shall not attempt to review these here, beyond mentioning that they are all relevant for the kinds of models that have been discussed.

4.2 Comparisons between geostatistical and MRF approaches

One aspect of spatial models which has not been very much explored is the connection between the geostatistical approches of sections 1 and 2, and the MRF models of the present section.

In multivariate normal cases, one way to characterize the difference is that geostatistical approaches work by specifying the covariance matrix of the observations, Σ say, in terms of unknown parameters θ. The MRF approach, through (28) , amounts to a parametric model for the inverse autocovariance matrix, $B = \Sigma^{-1}$. There is no obvious reason for preferring one to the other.

One issue is that of *marginalization* (the author thanks Julian Besag for drawing his attention to this issue). Suppose we have a random field defined for a continuous space variable. In order to fit it in with a MRF approach, we must first restrict the data to some form of lattice. For example, one approach which has been adopted in some agricultural or epidemiological contexts is to aggregate data by county, treating counties with common boundaries as neighbors in the MRF specification. But the question then arises: how much are the resulting joint distributions invariant to the arbitrary specifications of county boundaries?

As a concrete example of this problem, consider a model with data $x_1, ..., x_{n+1}$ corresponding to averages of some spatially distributed quantity over $n+1$ counties, and suppose a joint density $p_1(x_1, ..., x_{n+1})$ is specified. Now suppose some administrative authority decides to amalgamate the nth and $(n+1)$st counties, defining $x'_n = (x_n + x_{n+1})/2$. Let $p_2(x_1, ..., x_{n-1}, x'_n)$ denote the joint distribution under the new model. Logically, p_2 should be derived from p_1 through a marginalization condition of the form

$$p_2(x_1, ..., x_{n-1}, x'_n) = \int p_1(x_1, ..., x_{n-1}, x_n, 2x'_n - x_n) dx_n. \qquad (32)$$

The question then arises: do natural specifications of the models p_1 and p_2 satisfy relationships such as (32) ?

Although this is a very specific question it is meant to illustrate a general point, that there are natural consistency relations among probability distributions, and it may require some care to ensure that there are satisfied.

In the case of geostatistical models, the difficulty just described does not arise, because such models work by specifying the covariance between any pair of sites, and the covariance between two county averages is computed by an obvious integral of the pointwise covariance function over the two counties. Such an operation will always satisfy consistency relations such as (32) . However, for a typical MRF model, (32) will not be satisfied unless unusual care is taken in specifying the model. This appears to be an argument against the use of MRF models in cases where the lattice structure is not defined by the natural geometry of the system.

On the other hand, from other points of view, MRF models are more flexible — for example, we have already seen that they may be defined for binary data (through the autologistic model) and there are by now many models for count data with either marginal or conditional Poisson distributions, whereas all of our geostatistical discussion has been (implicity or explicitly) for normally

distributed data. For this reason, the MRF models potentially lead to a much richer class of models.

4.3 Use of MRFs in hierarchical models

One of the most rapid developments in recent years has been the use of MRFs as a component model in some hierarchical structure, this greatly increasing the scope of applications for such models. As an example, Best *et al.* (1998) considered extensions of the Clayton-Kaldor (1987) model of the form

$$Y_i \mid \mu_i \sim Pois\{E_i \exp(\mu_i)\},$$
$$\mu_i = x_i^T \beta + u_i + v_i, \tag{33}$$

in which Y_i is the count (of disease incidences, say) in a particular county or region i, E_i is the susceptible population and μ_i is a random intensity function; μ_i is specified through a regression model $x_i^T \beta$ and additional random errors u_i and v_i. In the model of Best *et al.*, v_i are independent random errors but u_i are spatially dependent and specified through a conditional autoregressive (or autonormal) model.

Diggle *et al.* (1998) proposed something similar but based on geostatistical models for μ_i. Wolpert and Ickstadt (1998) also considered a model with Poisson counts, but for them, the underlying field was assumed to have a special structure with gamma marginal distributions. Evidently, these are only a few examples of what is possible with this kind of structure, and there are many possibilities for extensions to other kinds of marginal or conditional distributions.

4.4 Environmental applications of MRF models

So far, the applications of MRF models in physical environmental modeling have been very limited, but an exception is the recent paper by Cressie *et al.* (1999). That paper applied both geostatistical and MRF approaches to the prediction of a particulate matter field based on 27 monitoring stations in the area of Pittsburgh, Pennsylvania. They also gave a more detailed discussion than we have here of the relative merits of the two approaches.

5 Spatial design of experiments

In this and the following section, we deal much more briefly with two other topics which are of major importance, but which space does not permit us to develop in more detail in the present review.

The issue of *spatial design of experiments* arises most commonly in developing a monitoring network. To take one example where the problem arises, the

U.S. Environmental Protection Agency is responsible for monitoring a large number of airborne and water-borne pollutants. There are legal requirements and political considerations in deciding where to place monitors but, in most cases, the Agency still has discretion over how many monitors to place in a particular city or region, and over the precise location of these monitors. For example, in a city such as Chicago, should there be a higher concentration of monitors near Lake Michigan (where, experience shows, there is a higher variability in atmospheric conditions due to local meteorological and lake-based effects), or should the monitors be evenly distributed over the suburban regions near the city? Another version of the problem (Oehlert, 1996) is when there is an existing network of monitors but, for cost-saving reasons, it is desired to close down a certain fraction of the network.

One formulation of the problem developed by Le and Zidek (1992) is to assume a relatively large but discrete set of potential sites of interest, divided into "gauged" sites (the ones where monitors are actually located) and "ungauged" sites. In broad terms, the problem then becomes to select the set of gauged sites so that the predictions at the ungauged sites are as accurate as possible. There are then two issues: (i) how to specify a suitable criterion for accuracy at a large set of ungauged sites, (ii) how to find designs which perform well under such a criterion.

Problem (i) is essentially the classical problem of optimal design of experiments, for which there are classical criteria such as D-optimality, E-optimality and so forth (see, e.g., Fedorov (1972) or Atkinson and Donev (1992)), or Bayesian approaches which are often formulated in terms of information-theoretic criteria, following Bernardo (1979).

Even when the criterion is well defined, the problem of selecting, for example, the best 99 out of 249 potential sites (Oehlert 1996) is a formidable combinatorial problem which defies exact solution. In practice, *ad hoc* addition-deletion rules have been developed. Other references include Brown *et al.* (1994), Oehlert (1993, 1995), Le *et al.* (1997) and Nychka and Saltzman (1998).

6 Spatial-temporal data

Spatial-temporal analysis is concerned with random fields of the form $X(s,t)$, where s is a location or site variable and t is time. (In contrast to previous sections, here we use t specifically to denote time.) A model for the random field then requires that we specify the joint distribution of $\{X(s_1,t_1),...,X(s_n,t_n)\}$ for any combination of space-time pairs $(s_1,t_1),...,(s_n,t_n)$. In the simplest case where we assume the field is Gaussian, this means spacifying the covariance between $X(s,t)$ and $X(s',t')$ for any s,t,s',t'.

The simplest models are the *separable models*, for which the spatial-temporal

covariance function factorizes as

$$\mathrm{Cov}\left\{X(s,t), X(s',t')\right\} = C(s,s')\gamma(t-t') \tag{34}$$

where $C(\cdot,\cdot)$ is a spatial covariance function and $\gamma(\cdot)$ is the covariance function of a stationary time series.

An early example of the application of (34) was the paper by Haslett and Raftery (1989), who used it to model the joint distribution of wind speeds at 12 stations in Ireland. In their model, spatial covariance was represented by a stationary, isotropic model with an exponential variogram (15), while the temporal covariance function they adopted was the fractional ARIMA process (Beran 1994), which incorporates long-range dependence in time.

Apart from the convenience of the mathematical representation, another advantage of separable models is their computational tractability. For example, Mardia and Goodall (1993) considered the case of a $m \times n$ data matrix $\mathbf{X} = (X_{ik})$, where X_{ik} is the value at the ith spatial location and the kth time point, with a model of form

$$\mathbf{X} \sim N(\mu, \Lambda \otimes \Gamma)$$

where $\Lambda = (\lambda_{ij})$ is a spatial covariance matrix and $\Gamma = (\gamma_{kl})$ is a temporal covariance matrix. [The $\otimes$ notation is interpreted to mean $\mathrm{Cov}(X_{ik}, X_{jl}) = \lambda_{ij}\gamma_{kl}$.] Assuming multivariate normality, the joint density of $\mathbf{X}$ may be written as

$$(2\pi)^{-mn/2}|\Lambda|^{-m/2}|\Gamma|^{-n/2}\exp\left[-\frac{1}{2}\,\mathrm{tr}\left\{\Lambda^{-1}(X-\mu)\Gamma^{-1}(X-\mu)^T\right\}\right]. \tag{35}$$

The importance of (35) is that it shows that one only needs to compute the determinant and inverse for the matrices Λ and Γ, and not for the $(mn)\times(mn)$ matrix $\Lambda \otimes \Gamma$. To do the latter directly would, of course, be both much slower and would consume far more computer storage space.

Nevertheless, despite the pragmatic advantages of separable models, it is now increasingly recognized that they are not a realistic assumption for much practical data. It is likely that much work over the next few years will be devoted to the development of new non-separable models for spatial-temporal processes. At the present time, the literature is scattered, with few coherent themes. A few recent developments are:

(i) Carroll *et al.* (1997) proposed a model for the spatial-temporal distribution of atmospheric ozone in the Houston area. By examining correlation vs. distance plots at different time lags, they proposed a specific parametric form for the spatial-temporal covariance function and fitted it through a cross-validation-type method, full maximum likelihood being infeasible for the data sizes they were considering. The model appeared to have good practical properties for the specific data to which it was applied, but the general form

of the covariance does not appear to be positive definite (Cressie 1997), and without this, the model is of limited general utility.

(ii) Jones and Zhang (1997) have proposed a class of continuous-parameter models derived from stochastic partial differential equations.

(iii) The most promising approach at the present time is based on generalizations of the Kalman filter, in which the set of spatial variables at each time is viewed as a random vector with dynamic equations for the temporal evolution. The approach is highly computationally intensive, but it is feasible within the structure of Bayesian hierarchical models, and has the additional advantage (e.g., in an atmospheric science context) that the dynamical equations can sometimes be suggested by the physics of the process being observed. Some representative papers on this approach are those by Wikle *et al.* (1998), Wikle and Cressie (1999).

7 Hierarchical models for spatial trends

In the remainder of this chapter, we outline some recent ideas for modeling of a spatially varying temporal trend. The canonical problem is suggested by global warming: empirical data show an increasing temperature trend at many points of the earth's surface, which some scientists interpret as evidence of an anthropogenically induced greenhouse effect. However, the trend is not the same at all places. For example, within the continental United States, the increasing temperature trend is greatest in the northern midwest states, while in other parts of the country, such as the south east, there has been little or no observed trend. This effect is clearly seen in individual time series at different spatial locations, but there is wide variability in estimated trends from one location to another which may simply be due to the statistical error in estimating the trends. Therefore, the problem arises of "smoothing" the trends available from individual time series, to obtain an overall picture of how the trend varies with space. A similar problem has been studied in connection with trends in atmospheric SO_2 levels across the United States (Holland *et al.* 2000), and in a rather different context, for the variability of particulate matter-based mortality across different cities (Dominici *et al.* 2000).

One plausible model is as follows. Suppose there is a linear spatially dependent trend, denoted $Z_1(s)$ for location s, for which

$$\mathrm{E}\{Z_1(s)\} = x(s)^T\beta, \quad \mathrm{Cov}\{Z_1(s), Z_1(s')\} = C(s, s'; \theta), \tag{36}$$

in which $x(s)$ is a known spatially dependent covariate vector, β an unknown vector of regression coefficients, and C a spatial covariance function depending on parameters θ. We assume that we cannot observe $Z_1(s)$ directly, but instead, for a fixed set of spatial locations $s = s_1, ..., s_m$, we observe a time series $\{Y(s, t_1), ..., Y(s, t_n)\}$, whose distribution depends on $Z_1(s)$ as well as

other unknown parameters which we shall denote by ϕ. This suggests a natural "hierarchical model" structure in which there is a top level of the hierarchy represented by the unknown parameters (β, θ, ϕ), a middle level represented by the unobserved process $Z_1(\cdot)$, and a bottom level represented by the observed data $\{Y(s_i, t_j),\ i = 1, ..., m,\ j = 1, ..., n\}$. The (Bayesian) specification of the model is completed by a prior distribution on (β, θ, ϕ), and the whole structure would then be analyzable by modern methods of hierarchical models analysis (see, e.g., Gilks *et al.*, 1996). However, for typical data sets with large numbers of spatial locations as well as many time points, such an approach, directly implemented, would be very time consuming.

We therefore propose an alternative approach which avoids the full complications of a hierarchical models analysis. Suppose, for each observed spatial location s_i, we calculate an estimate of $Z_1(s_i)$, which we denote $\tilde{Z}_1(s_i)$, based just on the time series $Y(s_i, t_j),\ j = 1, ..., n$. This may be based on any model appropriate for that time series. Since most statistical methods lead to approximately normal distributions of estimators in large samples, we may assume

$$\tilde{Z}_1(s_i) = Z_1(s_i) + \xi(s_i), \tag{37}$$

where $\{\xi(s_1), ..., \xi(s_m)\}$ is a zero-mean vector of errors such that

$$\mathrm{Cov}\{\xi(s_i), \xi(s_j)\} = w_{ij}, \tag{38}$$

$W = (w_{ij})$ being the error covariance matrix. Moreover, these errors, corresponding to measurement errors at individual stations, may be assumed independent of the true trend surface $Z_1(\cdot)$.

By combining (36)–(38), we have a model

$$\mathrm{E}\{\tilde{Z}_1(s_i)\} = x(s_i)^T\beta, \quad \mathrm{Cov}\{\tilde{Z}_1(s_i), \tilde{Z}_1(s_j)\} = C(s_i, s_j; \theta) + w_{ij}, \tag{39}$$

together with (approximate) joint normality.

We make one final simplifying assumption, which is to assume that the $\{w_{ij}\}$ in (38) are known. This is justified by the fact that, since these estimates arise from the statistical errors in individual time series, they can be characterized from standard error calculations in the individual time series estimations. At any rate, we should be able to approximate the w_{ij}s much more accurately than we could initially guess the covariances of the true Z_1 process, in other words, the C function in (39).

Thus we are led to a model of the form (39) in which $C(\cdot, \cdot; \theta)$ is a parametric covariance function depending on parameter vector θ, and $W = (w_{ij})$ is a known error covariance function.

We now give two examples of this approach. The first is again based on the Historical Climatological Network (see Section 3), from which we have calculated time trends in winter mean temperatures, over the period 1965–1996, for each of 184 stations, representing the data as the sum of a linear

trend and an AR(p) staionary time series and estimatoing the slope of the linear trend by maximum likelihood. This leads to an estimate $\tilde{Z}_1(s_i)$ of the temperature trend at each station, with an accompanying standard error. The squares of the standard errors are taken as values of the diagonal entries w_{ii}, while the off-diagonal entries, w_{ij} for $i \neq j$, are taken as 0. A homogeneous spatial model with Gaussian semivariogram function (16) is assumed for Z_1, and a regression model $x(s)^T\beta$ corresponding to a cubic polynomial function of s, after testing various alternative models both for the spatial covariance function and for the polynomial regression model. Kriging is then used to construct an estimate of the surface $Z_1(s)$. The resulting estimate (contour plot at the top, perspective plot at the bottom) is shown in Fig. 10. The result shows how the estimated trend varies across the country, with the largest trend around the great lakes region, consistent with the earlier description. Although we do not give error estimates here, it should be pointed out that the methodology does allow us to obtain approximate error bounds of the reconstructed surface.

The second example is based on Holland *et al.* (2000). In this paper, sulfur dioxide measurements at 35 locations in the eastern U.S. over the time period 1989–1995 were characterized as functions of seasonal trends, meteorology, and an overall additive linear trend. A generalized additive model (Hastie and Tibshirani, 1990), applied to the logarithms of weekly sulfur dioxide totals, was used to estimate the trend at each station, after adjusting for the seasonal and meteorological terms. In this example, instead of assuming the $W = (w_{ij})$ matrix in (39) is diagonal, w_{ij} is estimated for each (i, j) pair by a jackknife procedure. The model (39) is then fitted, again with a Gaussian semivariogram kernel (16). The resulting estimate of the trend surface is shown in Fig. 11.

The importance of this analysis, in the context of evaluating improved regional-scale air quality resulting from electric utility emission reductions, is that it allows the characterization of estimated trends in sulfur dioxide, not only at the monitoring stations themselves, but also on a regional basis. The trends can be compared to corresponding changes in sulfur dioxide emissions to evaluate the impact of reduced emissions. For the period, 1989–1995, reduced emissions levels from large electric utilities are similar to the estimates of regional trends.

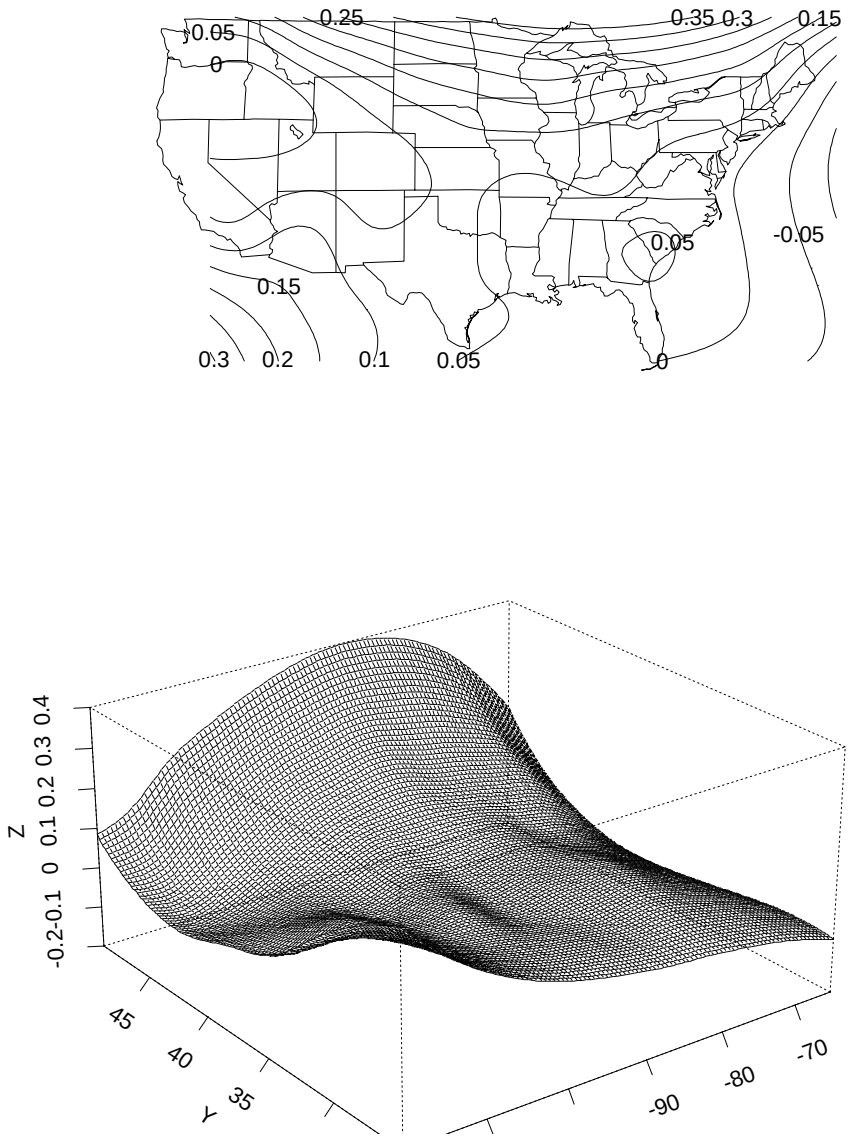

Fig. 10. Reconstructed trend surface for U.S. temperatures. Top picture: contour plot. Bottom picture: perspective plot. In the perspective plot, increasing values of Y correspond to increasing latitude ^{o}N, and increasing X to decreasing longitude (^{o}W).

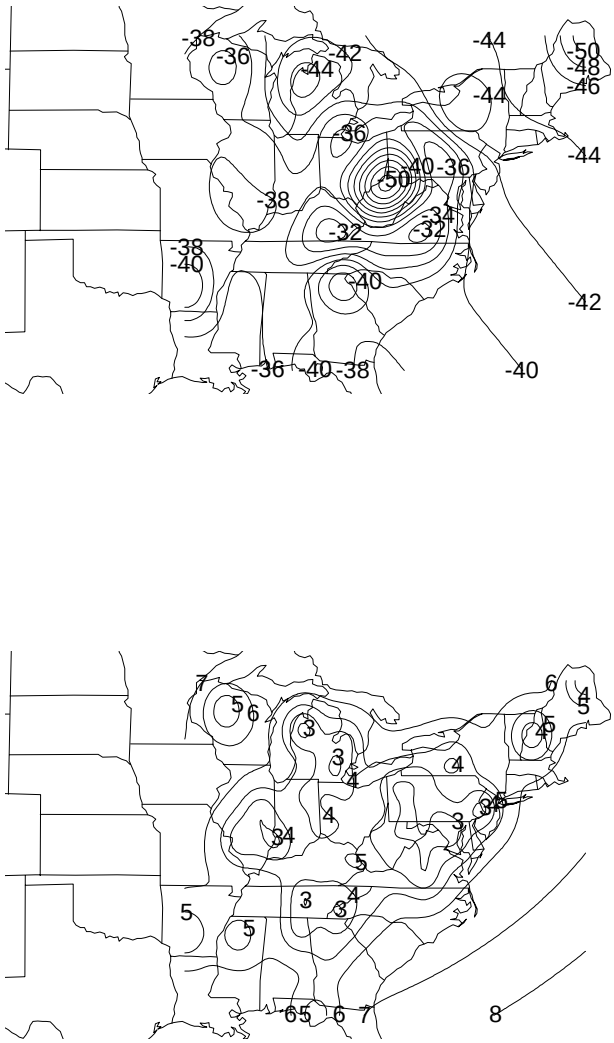

Fig. 11. Reconstructed trend surface for sulfur dioxide data (top plot) and prediction error variances (bottom plot).

8 Conclusions

In this chapter, I have attempted to give a broad overview of current themes in spatial statistics, though concentrating on the geostatistical approach, which remains the most widely applied method in environmental statistics. Modern methods of estimation, such as REML or Bayesian estimation, allow these processes to be estimated without some of the *ad hoc* features of earlier proposals, and the Bayesian procedures in particular have the advantage that when applied to the spatial prediction or kriging problem, they automatically allow for the uncertainty of the estimated model parameters, a deficiency of classical kriging. Extensions to nonstationary processes and spatial-temporal models are major themes of current research and may be expected to remain so for some time to come.

The final part of the chapter discussed a particular application of these techniques, to the estimation of spatially dependent trends. The method described in section 7 is intended to be fairly straightforward to apply as an extension of classical kriging, but here also there are possibilities for more general approaches, including fully Bayesian approaches.

References

Abramowitz, M. and Stegun, I.A. (1964), *Handbook of Mathematical Functions*. National Bureau of Standards, Washington D.C., reprinted by Dover, New York.

Atkinson, A.C. and Donev, A.N. (1992), *Optimum Experimental Designs*. Oxford University Press.

Beran, J. (1994), *Statistics for Long-Memory Processes*. Chapman and Hall, New York.

Bernardo, J.M. (1979), Expected information as expected utility. *Annals of Statistics* **7**, 686-690.

Besag, J. (1974), Spatial interaction and the statistical analysis of lattice systems. *J.R. Statist. Soc. B* **36**, 192–225.

Besag, J. (1975), Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195.

Besag, J. (1977), Efficiency of pseudolikelihood estimation for simple Gaussian field. *Biometrika* **64**, 616–618.

Besag, J., Green, P.J., Higdon, D. and Mengersen, K. (1995), Bayesian computations and stochastic systems. *Statistical Science* **10**, 1–66.

Besag, J. and Higdon, D. (1999), Bayesian analysis of agricultural field experiments (with discussion). *J.R. Statist. Soc. B* **61**, 691–746.

Besag, J. and Moran, P.A.P. (1975), On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika* **62**, 555–562.

Best, N.G., Arnold, R.A., Thomas, A., Waller, L.A. and Conlon, E.M. (1999), Bayesian models for spatially correlated disease and exposure data (with discussion). In *Bayesian Statistics 6*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press, pp. 131–156.

Brown, P.J., Le, N.D. and Zidek, J.V. (1994), Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics* **22**, 489–509.

Carroll, R.J., Chen R., George, E.I., Li, T.H., Newton, H.J., Schmiediche, H. and Wang, N. (1997), Ozone exposure and population density in Harris County, Texas. *J. Amer. Statist. Assoc.* **92**, 392–404.

Clayton, D.G. and Kaldor, J. (1987), Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrika* **43**, 671–681.

Clifford, P. (1990), Markov random fields in statistics. In *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley* (G.R. Grimmett and D.J.A. Welsh, editors). Oxford: Oxford University Press.

Comets, F. (1992), On consistency of a class of estimators for exponential families of Markov random fields on a lattice. *Annals of Statistics* **20**, 455–468.

Cressie, N. (1993), *Statistics for Spatial Data*, second edition. John Wiley, New York.

Cressie, N. (1997), Comment on Carroll *et al.* (1997). *J. Amer. Statist. Assoc.* **92**, 411–413.

Cressie, N., Kaiser, M.S., Daniels, M.J., Aldworth, J.W., Lee, J., Lahiri, S.N. and Cox, L.H. (1999), Spatial analysis of particulate matter in an urban environment. Preprint, Iowa State University.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998), Model-based geostatistics (with discussion). *Applied Statistics* **47**, 299–350.

Dominici, F., Samet, J.M. and Zeger, S.L. (2000), Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy (with discussion). *J.R. Statist. Soc. A* **163**, to appear.

Fedorov, V.V. (1972), *Theory of Optimal Experiments*. Academic Press, New York.

Geman, S. and Geman, D. (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geyer, C.J. and Thompson, E.A. (1992), Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J.R. Statist. Soc. B 54*, 657–699.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (eds.) (1996), *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Guttorp, P., Meiring, W. and Sampson, P. (1994), A space-time analysis of ground level ozone data. *Environmetrics* **5**, 241–254.

Haas, T.C. (1990), Lognormal and moving-window methods of estimating acid deposition. *J. Amer. Statist. Assoc.* **85**, 950–963.

Haas, T.C. (1995), Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *J. Amer. Statist. Assoc.* **90**, 1189–1199.

Handcock, M.S. and Stein, M. (1993), A Bayesian analysis of kriging. *Technometrics*, **35**, 403-410.

Harville, D.A. (1974), Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383-385.

Harville, D.A. and Jeske, D.R. (1992), Mean squared error of estimation or prediction under a general linear model. *J. Amer. Statist. Assoc.* **87**, 724-731.

Haslett, J. and Raftery, A.E. (1989), Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource. *Applied Statistics* **38**, 1-21.

Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models.* Chapman and Hall, London.

Holland, D.M., De Oliveira, V., Cox, L.H. and Smith, R.L. (2000), Estimation of regional trends in sulfur dioxide over the eastern United States. *Environmetrics*, to appear.

Holland, D., Saltzman, N., Cox, L.H. and Nychka, D. (1999), Spatial prediction of sulfur dioxide in the eastern United States. In *geoENV II — Geostatistics for Environmental Applications*, eds. Gómez-Hernández, J., Soares, A. and Froidevaux, R., Kluwer, Dordrecht, 65–76.

Jones, R.H. and Zhang, Y. (1997), Models for continuous stationary spatial-temporal processes. In *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions*, edited by T.G. Gregoire *et al.*. *Lecture Notes in Statistics* **122**, Springer Verlag, New York, pp. 289–298.

Kitanidis, P.K. (1983), Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research* **19**, 909-921.

Le, N.D. and Zidek, J.V. (1992), Interpolation with uncertain spatial covariances: A Bayesian alternative to kriging. *Journal of Multivariate Analysis* **43**, 351–374.

Le, N.D., Sun, W. and Zidek, J.V. (1997), Bayesian multivariate spatial interpolation with data missing by design. *J.R. Statist. Soc. B* **59**, 501–510.

Mardia, K.V. and Goodall, C.R. (1993), Spatial-temporal analysis of mulitvariate environmental monitoring data. In *Multivariate Environmental Statistics*, eds. G.P. Patil and C.R. Rao, Elsevier Science Publishers, pp. 347–386.

Mardia, K.V. and Marshall, R.J. (1984), Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135-146.

Nychka, D. and Saltzman, N. (1998), Design of Air Quality Monitoring Networks. Chapter 4 of *Case Studies in Environmental Statistics*, edited by

D. Nychka, W. Piegorsch and L.H. Cox. Springer Lecture Notes in Statistics, number 132, Springer Verlag, New York, pp. 51–76.

Oehlert, G.W. (1993), Regional trends in sulfate wet deposition. *Journal of the American Statistical Association* **88**, 390-399.

Oehlert, G.W. (1995), The ability of wet decomposition networks to detect temporal trends. *Environmetrics* **6**, 327–339.

Oehlert, G.W. (1996), Optimal shrinking of a wet decomposition network. *Atmospheric Environment* **30**, 1347–1357.

Penttinen, A. (1984), Modelling interaction in spatial point patterns: parametric estimation by the maximum likelihood method. *Jy. Stud. Comput. Sci. Econ. Statist.* **7**.

Propp, J.G. and Wilson, D.B. (1996), Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252.

Ripley, B.D. (1981), *Spatial Statistics.* Wiley, New York.

Ripley, B.D. (1988), *Statistical Inference for Spatial Processes.* Cambridge University press, Cambridge, U.K.

Sampson, P.D. and Guttorp, P. (1992), Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87**, 108-119.

Smith, R.L. (1996), Estimating nonstationary spatial correlations. Unpublished; University of North Carolina, Chapel Hill. Available at

www.unc.edu/depts/statistics/postscript/rs/nonstationary.ps.

Swendsen, R.H. and Wang, J.-S. (1987), Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58**, 86–88.

Waller, L.A., Carlin, B.P., Xia, H. and Gelfand, A.E. (1997), Hierarchical spatio-temproal mapping of disease rates. *J. Amer. Statist. Assoc.* **92**, 607–617.

Wikle, C., Berliner, L.M. and Cressie, N. (1998), Hierarchical Bayesian space-time analysis. *Journal of Environmental and Ecological Statistics* **5**, 117–154.

Wikle, C. and Cressie, N. (1999), A dimension reduction approach to space time Kalman filtering. *Biometrika* **86**, 815–829.

Wolpert, R.L. and Ickstadt, K. (1998), Poisson-gamma random field models for spatial statistics. *Biometrika* **85**, 251–267.

Zimmerman, D.L. and Cressie, N. (1992), Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Statist. Math.* **44**, 27-43.