ESTIMATING NONSTATIONARY SPATIAL CORRELATIONS

by

Richard L. Smith [1]

Cambridge University

and

University of North Carolina at Chapel Hill

Preliminary Version: June 13 1996

Summary

The paper is concerned with the estimation of a spatial correlation structure under circumstances when the usual assumptions of stationarity and isotropy do not apply. An ingenious approach due to Sampson and Guttorp is based on a nonlinear transformation of the sampling space into an alternative space within which the spatial structure is stationary and isotropic. However the actual algorithm devised by Sampson and Guttorp is complicated and has a number of *ad hoc* features. In this paper we consider alternative methods based on parametric maximum likelihood fits, using a radial basis function representation of the nonlinear map. A key part of the fitting procedure is model selection, or equivalently, reduction in dimensionality by selection of a subset of radial basis functions. The methodology is illustrated with two examples, one based on tropospheric ozone and the other on U.S. climate data. However a number of cautions are noted: there is no guarantee of uniqueness of the estimates and the evidence that more complicated models result in improved spatial predictions is, at best, inconclusive.

1. INTRODUCTION

Environmental data are often collected on an irregular grid of spatial locations, and it is important to understand the spatial covariance structure. Besides the classical geostatistical problem of prediction at unmonitored sites ("kriging"), there are also important applications concerned with trend estimation and spatial design. The problem of trend

[1] Current address (to July 31 1996): Statistical Laboratory, 16 Mill Lane, Cambridge CB2 1SB, U.K. Thereafter: Department of Statistics, University of North Carolina, Chapel Hill, N.C. 27599-3260, U.S.A. Email rsmith@statslab.cam.ac.uk or rs@stat.unc.edu. Part of this research was carried out at the National Institute of Statistical Sciences (Research Triangle Park, North Carolina) where it was supported by the U.S. Environmental Protection Agency under Cooperative Agreement #CR819638-01-0.

estimation concerns how to measure, or to determine the significance of, a trend in the presence of spatial and/or temporal correlations in the data. For example, this problem frequently arises in the analysis of global climate change (Handcock and Wallis, 1994). Spatial design problems arise in deciding, for example, where to locate monitoring stations so as to maximize the information obtained about a pollutant over a broad geographical area. This problem has been much studied in recent years, see e.g. Oehlert (1995). All of these problems require, in the first instance, a model for spatial covariances.

We consider a spatial process $\{Z(s),\ s \in D\}$, where $D \subseteq \mathbb{R}^d$ is a domain of spatial locations. Usually $d = 2$ or 3 (in this paper, $d = 2$ always). Spatial dependence is usually characterized in terms of either the *covariance function*

$$C(s_1, s_2) = \text{ Cov } \{Z(s_1), Z(s_2)\}, \quad s_1, s_2 \in D,$$

or the *dispersion*

$$D(s_1, s_2) = \text{ Var } \{Z(s_1) - Z(s_2)\}, \quad s_1, s_2 \in D.$$

Much of the literature is concerned with processes which satisfy some or all of

(i) *intrinsic stationarity*: $D(s_1, s_2)$ depends on s_1 and s_2 only through the (vector) difference $s_1 - s_2$,

(ii) *stationarity*: $C(s_1, s_2)$ depends on s_1 and s_2 only through $s_1 - s_2$ (this implies intrinsic stationarity, but not conversely),

(iii) *isotropy*: $D(s_1, s_2)$ or $C(s_1, s_2)$ depends only on $||s_1 - s_2||$ (the Euclidean norm of $s_1 - s_2$ or, equivalently, the Euclidean distance between the locations s_1 and s_2). In this case we often write $D(s_1, s_2) = 2\gamma_0(||s_1 - s_2||)$, where γ_0 is an isotropic semivariogram function.

When all of (i)–(iii) hold we shall call the process *homogeneous*.

Classical geostatistics is concerned primarily with homogeneous processes for which, by now, a very extensive literature exists — see, for example, the comprehensive review by Cressie (1993). Until recently, however, not much was known about the modeling of inhomogeneous processes. One old approach (Journel and Huijbregts 1978) for stationary, non-isotropic processes is to write the dispersion in the form

$$D(s_1, s_2) = 2\gamma_0(||A_0(s_1 - s_2)||)$$

or by extension

$$D(s_1, s_2) = 2 \sum_{j=0}^{J-1} \gamma_j(||A_j(s_1 - s_2)||). \tag{1.1}$$

Here $A_0, A_1, ...$, are arbitrary matrices and $\gamma_0, \gamma_1, ...$, isotropic semivariogram functions. However, this is still quite a restrictive class of models.

A much more radical extension has been proposed in a series of papers by Sampson and Guttorp — see, in particular, Sampson and Guttorp (1992). They considered models of the form

$$D(s_1, s_2) = 2\gamma_0(f(s_1), f(s_2)) \tag{1.2}$$

with γ_0 again an isotropic semivariogram and f a smooth nonlinear map from $\mathbb{R}^d$ to $\mathbb{R}^{d'}$. In principle one may permit $d' \neq d$ though in most of the Sampson-Guttorp work it is assumed that $d' = d$ and we shall continue to assume that here. The idea behind (1.2) is that the map f takes the coordinates from the real, geographical or "G" space, into an alternative dispersion or "D" space in which the process is homogeneous. This approach may not be universally applicable to inhomogeneous processes, and in Section 5 of the present paper we shall see some limitations to it, but the model represents such a significant departure from previous practice, such as (1.1), that it has rightly attracted a lot of attention. One early illustration of the effectiveness of spatial distortion, in this case applied to the map of Ireland, was given by Lewis (1989). Other examples discussed by Sampson and Guttorp (1992), Guttorp, Sampson and Newman (1992) and Mardia and Goodall (1993) include solar radiation data from a number of sites near Vancouver, B.C., and data sets drawn from the UAPSP (acid rain) study. In sections 4 and 5 we shall discuss applications to ground-level ozone data and to climatology. Thus the potential applications of (1.2) appear to be very broad.

The precise methodology used by Sampson and Guttorp to fit (1.2), however, contained a number of rather *ad hoc* features. Briefly, it consists of three stages:

(a) A mapping of the n sampling points from the G space into the D space is found to minimize a stress criterion

$$\min_{\delta} \frac{\sum_{i<j}\{\delta(d_{ij}) - h_{ij}\}^2}{\sum_{i<j} h_{ij}^2}$$

where d_{ij} is the observed dispersion between sites i and j, h_{ij} is the distance between sites i and j in D space and the minimization is taken over all monotonically increasing functions δ. This formulation of the problem permits it to be solved by a multidimensional scaling (MDS) algorithm.

(b) The mapping of the N sampling points is then extended to a *smooth* function from the entire G space into the D space, using a representation based on thin plate splines.

(c) The function δ is replaced by a smooth function g (so $d_{ij} \approx g(h_{ij})$), which satisfies the positive definiteness condition required for g to be the variogram of a homogeneous process. For this purpose, Sampson and Guttorp used a very general representation of g as a mixture of Gaussian-type variograms.

The central theme of the present paper is to propose an alternative, likelihood-based, approach to fitting the model (1.2). There are a number of reasons why this might be considered desirable:

(i) The maximum likelihood approach to spatial processes was first proposed by Mardia and Marshall (1984) and has been shown to be computationally feasible for data sets consisting of as many as several hundred sites. This has not been unchallenged, since Warnes and Ripley (1987) argued that the likelihood may be highly multimodal even in the case of a very simple homogeneous model for the spatial correlations, and illustrated this on a particular geostatistical data set. However, Mardia and Watkins (1989) analysed the same data set, and claimed that the likelihood for this example is in fact unimodal — a claim which, based on independent computations, I believe to be correct. Maximum likelihood approaches have the potential advantages (though these will not be considered in the present paper) of being automatically extendable to models including a spatially varying mean and to spatial-temporal processes. Nevertheless, multimodality *is* an issue in fitting complicated spatial models, as we shall see.

(ii) In recent years, stimulated by Handcock and Stein (1993) and several papers by Zidek and co-authors, attention has turned towards Bayesian approaches to spatial data analysis. For example, Brown, Le and Zidek (1994) considered a model of the form

$$Z = \begin{pmatrix} Z^{(1)} \\ Z^{(2)} \end{pmatrix},$$

in which $Z^{(2)}$ represents a set of "gauged" sites at which data are available, and $Z^{(1)}$ a set of "ungauged" sites for which predictions are sought. Their hierarchical model was of the form

$$\begin{aligned} Z|X, B, \Sigma &\sim N(BX, \Sigma), \\ B|B_0, \Sigma, F &\sim N(B_0, \Sigma \otimes F^{-1}), \\ \Sigma|\Psi, m &\sim W^{-1}(\Psi, m), \end{aligned}$$

where N and W^{-1} represent normal and inverse Wishart distributions of the appropriate dimensions, X is a matrix of known regressors, and B_0, F and Ψ respresent parameters of the prior distribution. A key component of this, essential for any meaningful extrapolation from the gauged to the ungauged sites, is a representation for Ψ, which can be thought of as a prior guess for Σ. To estimate Ψ for an inhomogeneous process, Brown, Le and Zidek in effect used the Sampson-Guttorp technique — but, not being likelihood-based, this necessarily takes us outside a formal hierarchical Bayesian approach. Although the approach taken in the present paper is non-Bayesian, by developing representations for the likelihood function, it also provides a framework for Bayesian analysis of spatial models.

(iii) The Sampson-Guttorp approach implies some restrictions on the models considered. In particular, by using MDS to model the locations so that increasing distances correspond to increasing dispersions, the possibility that g may be non-monotone is excluded. Our approach does allow non-monotonic g, a feature that turns out to be important in Section 4.

Apart from Sampson and Guttorp (1992), a number of other approaches to inhomogeneous spatial process have been given. Loader and Switzer (1992) used generalized additive models to develop a broad class of models for an inhomogeneous spatial trend surface, but still with isotropic covariances. This approach may well be adequate for many inhomogeneous processes, but not when there is clear evidence of inhomogeneity in the covariances. The ozone data of Section 3 provides a clear example of this. Much closer to our approach is Mardia and Goodall (1993), who also used a maximum likelihood approach for a class of models very similar to those discussed in section 2. In fact, they considered some more general features — Box-Cox transformations on the observations, multivariate data, and spatial-temporal models. However, they did not study the model selection aspects of the problem, which are an important part of the methodology developed here.

An outline of the remainder of the paper is as follows. Section 2 presents the basic approach and discusses its computational aspects. Section 3 discussses model selection, in particular the choice of centers for the radial basis function representation of f. Sections 4 and 5 present two examples, concerned with ozone and climate data respectively. Section 6 presents a simulation experiment designed to assess the usefulness of the methodology for spatial prediction — this contains some cautions about its applicability. Finally section 7 presents conclusions and some indications of future work to be done.

2. MAXIMUM LIKELIHOOD ESTIMATION

The main components of our model are as follows:

1. We have N replications $Z_1, ..., Z_N$ of a spatial field observed at each of n sites (thus $Z_k = (Z_k(s_1), ..., Z_k(s_n))$, where $s_1, ..., s_n$ are the sampling sites, for $k = 1, 2, ..., N$). These are assumed independent with

$$Z_k \sim N_n(\mu, \Sigma), \tag{2.1}$$

N_n denoting the n-dimensional normal distribution, μ an arbitrary n-vector of means and Σ an $n \times n$ covariance matrix.

2. We assume $\sigma_{ij} =$ Cov $\{Z_k(s_i), Z_k(s_j)\}$ of the form

$$\sigma_{ij} = C_0(f(s_i), f(s_j)) \tag{2.2}$$

where C_0 is a homogeneous covariance function and f is represented by a linear combination of radial basis functions ((2.5) below).

3. For the initial analysis we assume C_0 has the Matérn structure

$$C_0(t) = \frac{1}{2^{\theta_2 - 1}\Gamma(\theta_2)} \left(\frac{2\sqrt{\theta_2}t}{\theta_1} \right)^{\theta_2} \mathcal{K}_{\theta_2} \left(\frac{2\sqrt{\theta_2}t}{\theta_1} \right). \tag{2.3}$$

Here $\theta_1 > 0$ is the spatial scale parameter and $\theta_2 > 0$ is a shape parameter. The function $\Gamma(\cdot)$ is the usual gamma function while $\mathcal{K}_{\theta_2}$ is the modified Bessel function of the third kind of order θ_2 (Abramowitz and Stegun 1964, Chapter 9). This form was used by Handcock and Wallis (1994) for their analysis of climate data and it seems to be very widely applicable as a simple parametric form for spatial correlations. Neveretheless it is not universally appropriate and we shall propose a more general representation for two-dimensional isotropic covariances in (2.12) below.

One simplification adopted in the present paper is that we do not consider in any detail the estimation of μ in (2.1), but focus all our attention on Σ. Thus if we simply replace μ by the vector of sample means, the negative log likelihood based on $Z_1, ..., Z_N$ reduces to

$$L = \frac{N}{2} \log |\Sigma| + \frac{N-1}{2} \text{ tr } \left(\Sigma^{-1} \hat{\Sigma} \right) \tag{2.4}$$

where $\hat{\Sigma}$ is the usual $n \times n$ sample covariance matrix. In fact, we simplify the analysis further by focussing on the *correlation* matrix. Thus, $\hat{\Sigma}$ is the sample correlation matrix and Σ the correlation matrix determined by the model. From now on, all the analysis is based on correlation matrices via (2.4).

If f represents a univariate function of spatial coordinates $s_i = (x_i, y_i)$, then a familiar way to represent f nonparametrically is in terms of thin-plate splines (see, for example, Green and Silverman 1994, Chapter 7). Typically a function f is chosen to pass through a finite number of data points $f_i = f(x_i, y_i)$ $(i = 1, ..., n)$, to minimize the bending energy

$$J(f) = \int \int_{\mathbb{R}^2} \left\{ \left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right\} dx dy.$$

The solution to this problem may be represented in the form (Green and Silverman 1994, page 142)

$$f(x, y) = a + bx + cy + \sum_{i=1}^{n} \delta_i \eta_i(x, y) \tag{2.5}$$

where

$$\sum \delta_i = \sum \delta_i x_i = \sum \delta_i y_i = 0 \tag{2.6}$$

and

$$\eta_i(x, y) = r^2 \log r, \qquad r = \{(x - x_i)^2 + (y - y_i)^2\}^{\frac{1}{2}}. \tag{2.7}$$

Thus (2.5) represents f as a sum of linear terms and n *radial basis functions* η_i with *centers* at the observed data points (x_i, y_i). The constraints (2.6) ensure that the problem does not become overdetermined.

An *interpolating spline* is a function of the form (2.5)–(2.7) which satisfies $f(x_i, y_i) = f_i$ at each of the n data points. In statistics, however, one is usually more interested in a *smoothing spline*, which is typically presented as a solution to the problem of minimizing

$$S(f) = \sum \{f_i - f(x_i, y_i)\}^2 + \alpha J(f) \tag{2.8}$$

where $\alpha > 0$ is a smoothing parameter.

Although the exact solution to (2.8) is computable (Green and Silverman 1994, page 147–148), in practice we do not usually have an *a priori* fixed value of α and an alternative approach is simply to restrict the representation (2.5) to a subset of radial basis functions. Thus we assume

$$\delta_i = 0 \quad \text{for } i \notin \{i_1, ..., i_m\} \tag{2.9}$$

where $i_1, ..., i_m$ are some subset of indices to be determined. This approach is similar to the way radial basis functions have been used in non-linear time series analysis (Casdagli 1989, Smith 1993, Judd and Mees 1995), where they are an alternative to neural net representations.

In the present context f is a bivariate function, so we apply the RBF approach to each of its two components, $f^{(1)}$ and $f^{(2)}$ say. There is a potential difficulty with this in that a function constructed in this way may not be bijective. Non-bijective functions are a problem in the Sampson-Guttorp approach because they correspond to a mapping which folds over itself, which in most contexts seems counterintuitive. The difficulty is noted by Sampson and Guttorp (1992) who suggest that, in most cases, the problem of folding can be avoided by choosing a sufficiently smooth map, which is equivalent to keeping m, the number of active RBFs, fairly small.

Some further simplification is possible. First, the constant a in (2.5) is unnecessary — this is so because the resulting covariance functions depend only on *differences* between coordinates in the D space and are therefore unaffacted by locations shifts in D space. So we set $a = 0$. Second, in the case $m = 0$, the model is invariant under orthogonal rotations. This suggests that, in the case $m > 0$ as well, we simplify the parametrization to

$$\begin{aligned} f^{(1)}(x,y) &= b_1^2 x + \rho b_1 b_2 y + \sum_1^n \delta_i^{(1)} \eta_i(x,y), \\ f^{(2)}(x,y) &= \rho b_1 b_2 x + b_2^2 y + \sum_1^n \delta_i^{(2)} \eta_i(x,y), \end{aligned} \tag{2.10}$$

where $b_1 > 0,\ b_2 > 0,\ \rho \in \mathbb{R}$, and each of the sequences $\{\delta_i^{(1)},\ i = 1, ..., n\}$ and $\{\delta_i^{(2)},\ i = 1, ..., n\}$ satisfy the constraints (2.6), (2.9). Finally we note that with f still permitting arbitrary scale changes, we may without loss of generality set $\theta_1 = 1$ in (2.3). Thus, whenever $m \geq 3$, the final model has $2m-2$ free parameters $b_1,\ b_2,\ \rho,\ \theta_2,\ \delta_{i_1}^{(1)},\ \delta_{i_1}^{(2)}, ...,\ \delta_{i_{m-3}}^{(1)},\ \delta_{i_{m-3}}^{(2)}$.

The model (2.3) does not allow for a *nugget effect*. This could be permitted, for example, by allowing the value for $C_0(0)$ to be greater than the limiting $t \to 0$ value obtained from (2.3). In many applications, no nugget effect is observed with the Matérn covariance structure, but with the climate data of Section 5, it turns out that such an effect is needed. A much broader extension is to abandon the parametric form entirely and

to represent C_0 nonparametrically. A general representation for a d-dimensional isotropic covariance function (Cressie 1993, page 85) is given by

$$C_0(h) = \int_0^\infty Y_d(wh)\Phi(dw)$$

where $\Phi(\cdot)$ is a general positive measure on $[0, \infty)$ and

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

where J_v is the modified Bessel function of order v. In particular, when $d = 2$ this reduces to

$$C_0(h) = \int_0^\infty J_0(wh)\Phi(dw). \tag{2.11}$$

In practice the measure Φ may be assumed concentrated on a finite number of atoms, so (2.11) reduces to

$$C_0(h) = \sum_{c=1}^{C} \phi_c J_0(w_c h) \tag{2.12}$$

in terms of $2C$ parameters $\phi_1, w_1, ..., \phi_C, w_C$. Sampson and Guttorp (1992) used a similar representation but with $J_0(t)$ replaced by the Gaussian-type kernel e^{-t^2}. This is derived from the slightly less logical requirement that the function C_0 should be a positive definite isotropic covariance function in all dimensions simultaneously, rather than just in the specific dimension d in which we happen to be working. The practical importance of this is that the Sampson-Guttorp representation forces the covariance function to be monotone, whereas ours does not.

Calculation of the maximum likelihood estimates has been carried out using the Cholesky decomposition (Healy 19xx) to assist in evaluating $|\Sigma|$ and Σ^{-1}, followed by the quasi-Newton routine DFPMIN of Press *et al.* (1986) to minimize (2.4). This routine assumes that first-order derivatives of L are available, but these were calculated through a simple differencing approximation. The Matérn covariance function (2.3) was computed using a Fortran routine kindly supplied by Dr. M. Handcock, whilst J_0 in (2.12) was calculating via a combination of power series for small argument and asymptotic expansions for large argument (Abramowitz and Stegun 1964).

3. MODEL SELECTION

An important feature of the approach being developed in this paper is that only a subset of centers, represented by the indices $i_1, ..., i_m$ in (2.9), is included in the model. This is contrast to Mardia and Goodall (1993), who implicitly assumed that all the centers are included. Using all the centers leads to intractable computational problems when the

number of centers is large, and might also be expected to result in badly overfitted models. Therefore, we must limit the number of centers included, but the computational complexity of the problem rules out any attempt at an exhaustive search over subsets.

To simplify this problem, the n centers are first arranged in order, with indices $i_1, ..., i_n$. The problem then reduces to the selection of m, the number of centers to be included in the model. To determine the order of centers, different approaches have been adopted for the two examples to be considered later in the paper. With the ozone example of section 4, the total number of centers is comparatively small ($n = 21$), and this makes it possible to proceed somewhat interactively. The ideal would be to introduce the centers in an order which maximizes the log likelihood at each stage of the fitting proceduce. In practice even this is not easy to determine, but some trial and error along these lines did precede the actual choice of order which is used for the discussion in Section 4. With the climate example of Section 5, the size of the problem ($n = 138$) seems to preclude any attempt to determine an optimal ordering. In this case, the order was determined on purely geographical grounds, so as to achieve a good spread of stations at each stage. However, this part of the modeling procedure is admittedly somewhat arbitrary and I make no claim to any kind of optimality.

A more detailed analysis is possible for the choice of m, the number of centers to include. One approach is via the maximized log likelihoods, with either a sequence of likelihood ratio tests or some automatic model selection criteerion such as AIC. This tends to result in a large value of m being chosen, and there is evidence that models chosen in this way perform badly from a predictive viewpoint (Section 6). As an alternative, I propose a method based on cross-validation.

The traditional approach to cross-validation is based on leaving out one observation at a time, in order to predict it by re-fitting the model to the remaining observations. In the present context, that would mean leaving out one station at a time. However, with so much effort needed to calculate the maximum likelihood estimators, such an approach would be very time consuming. As a compromise between statistical and computational efficiency, I adopt an approach based on leaving out one quarter of the observations as each stage.

Suppose we write a typical data vector in the form

$$Z = \begin{pmatrix} Z^{(1)} \\ Z^{(2)} \end{pmatrix}$$

where $Z^{(1)}$ represents the approximately one quarter of the observations omitted, and $Z^{(2)}$ the remainder. We partition both the fitted correlation matrix Σ and the sample correlation matrix $\hat{\Sigma}$ in the obvious way,

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{pmatrix}.$$

The model is re-fitted using just the $Z^{(2)}$ components, and used to predict $Z^{(1)}$. Since we are not attempting to model the station means, we assume without loss of generality that they are all 0. The optimal predictor of $Z^{(1)}$ is then $\hat{Z}^{(1)} = AZ^{(2)}$, where $A = \Sigma_{12}\Sigma_{22}^{-1}$. The mean squared prediction error is given by

$$\begin{aligned} &\mathrm{E}\left\{(Z^{(1)} - \hat{Z}^{(1)})^T(Z^{(1)} - \hat{Z}^{(1)})\right\} \\ &= \mathrm{E}\left[\ \mathrm{tr}\ \left\{Z^{(1)}Z^{(1)T} - 2AZ^{(2)}Z^{(1)T} + AZ^{(2)}Z^{(2)T}A^T\right\}\right]. \end{aligned} \tag{3.1}$$

If we average (3.1) over the N data points, a sample-based estimate becomes

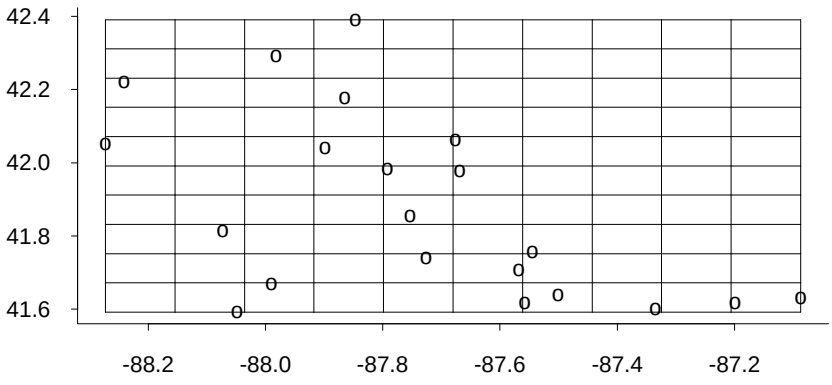
$$\mathrm{tr}\ \left\{\hat{\Sigma}_{11} - 2\Sigma_{12}\Sigma_{22}^{-1}\hat{\Sigma}_{21} + \Sigma_{12}\Sigma_{22}^{-1}\hat{\Sigma}_{22}\Sigma_{22}^{-1}\Sigma_{21}\right\}. \tag{3.2}$$

This calculation is repeated four times, with a different quarter of the stations omitted on each occasion. Finally, the four cross-validation scores obtained from (3.2) are added, to obtain an overall CV score for the model.

4. OZONE EXAMPLE

This example arose during the course of a larger study (Nychka and Royle, 1996) into the design of a monitoring network for ground-level ozone. Twenty-one monitoring stations from the greater Chicago area are shown in Fig. 4.1. The city of Chicago is roughly in the middle of the picture and the blank area to the upper right is part of Lake Michigan; there are no monitoring stations on the lake. We shall not attempt to review the network design problem itself, except to remark that it requires accurate modeling of spatial covariances. Henceforth we concentrate on this feature. The data consisted of a sample covariance matrix based on 89 vector observations, which we assume to be independent. As previously discussed, the analysis will be simplified by ignoring any variation in the station variances and focussing exclusively on the sample correlation matrix.

Fig. 4.1: Ozone Stations



As discussed in Section 3, a critical part of the analysis is to arrange the stations in order, and the order that has been adopted in this example has been determined by a certain amount of trial and error, with the intention that centers which make a large contribution to the model are introduced early in the analysis. I shall not attempt to describe this in any detail, however, and proceed immediately to the selection of m, the number of centers to be included in the model.

Table 4.1 lists the 19 models, starting with $m = 3$ (for which all the $\delta_i^{(1)}$ and $\delta_i^{(2)}$ coefficients in (2.10) are 0) and proceeding up to the full model $m = 21$. The table shows both the minimum L values and the CV scores. All the models are based on the Matérn form of covariance (2.3), with $2m - 2$ independent parameters. It can be seen that a likelihood ratio test or AIC approach would lead to a large value of m being chosen, such as $m = 19$. However, the approach based on minimizing the CV score leads to $m = 8$. This only just beats the models with $m = 3$, a linear transformation from G to D space!

Number of stations	Minimium L	CV score
0	598.7	3.34
4	649.3	3.35
5	672.3	3.47
6	689.2	3.49
7	701.3	3.53
8	745.3	3.33
9	753.7	3.44
10	754.3	3.44
11	765.7	3.66
12	772.1	3.74
13	772.3	4.01
14	777.9	4.33
15	782.2	5.22
16	793.0	7.68
17	802.0	6.44
18	805.6	3.92
19	813.6	7.10
20	813.9	5.32
21	815.8	4.33

Table 4.1: Minimum L values and CV scores for a sequence of models, ozone data.

A plot of the CV scores is shown in Fig. 4.2. We might expect these to decrease steadily for the first few values of m, to reach a minimum, and then to increase again. This is far from the observed form of the plot, a situation which may be due to the irregular spatial distribution of the stations combined with the strong nonlinearity of the optimization problem. However, it is clear that the CV scores in the right-hand part of

the plot (say, for $m > 12$) are substantially larger than those in the left-hand part, which indicates that we should not allow m to be too large. The subsequent discussion is based on $m = 8$.

Fig. 4.2: CV Scores for Ozone Data

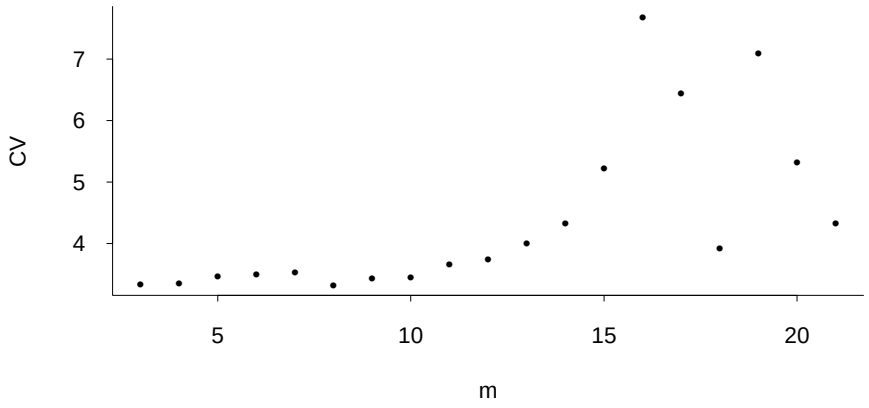
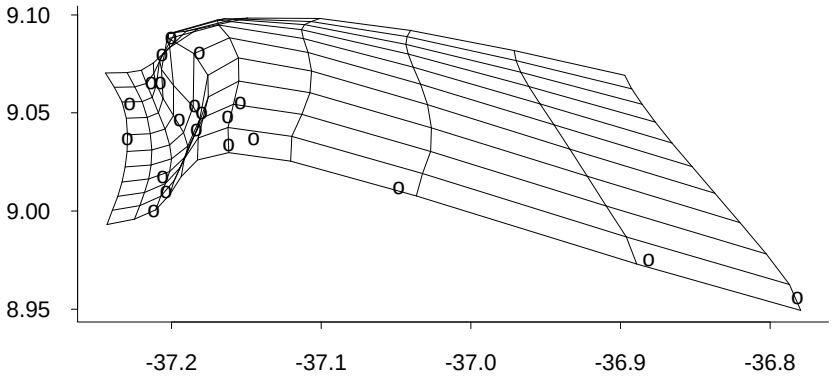


Fig. 4.3: D-space for Ozone Data



The D space under this transformation is shown in Fig. 4.3. The striking feature of this plot is that the three stations in the lower right-hand corner of Fig. 4.1 have been pulled a considerable distance from the other 18, which reflects the fact that the spatial correlations between the two groups, although still positive, are much smaller than those within the larger group. The explanation may lie in the distribution of sources. Ozone in

Chicago itself, and in the suburbs to the north and west, is caused primarily by traffic, whereas there are a number of industrial sources in the neighborhood of Gary, IN, where the three discrepant monitors are located.

Fig. 4.4: Semivariogram Plots in G Space

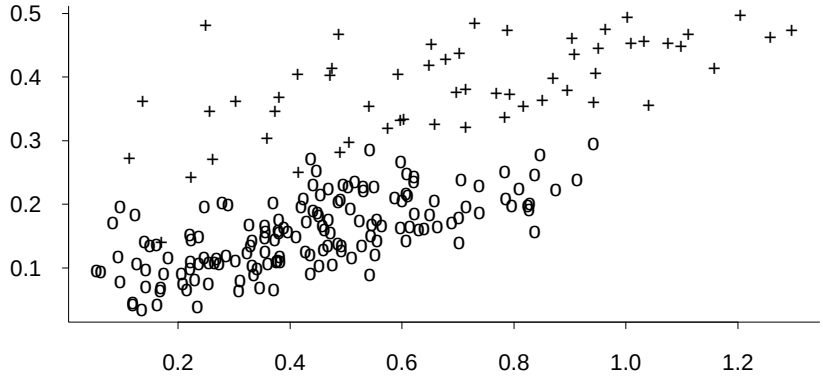
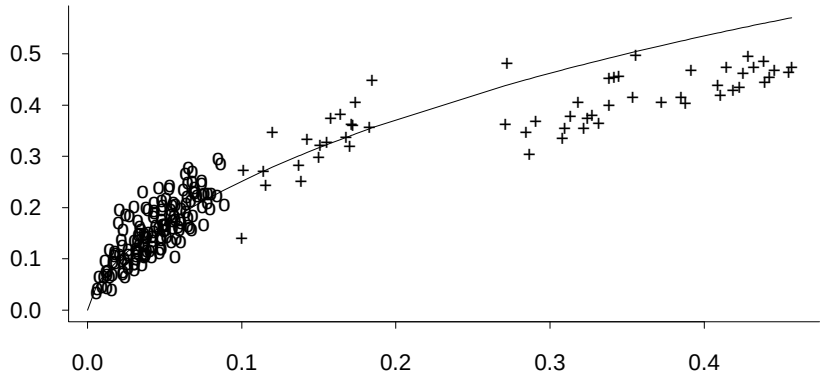


Fig. 4.4 shows the semivariogram plot in G space, and Fig. 4.5 the same plot in D space, together with the fitted Matérn curve. The variability of the plot is much reduced by transforming to D space and fits the Matérn curve reasonable well. Each pair of stations is represented by one point on the plot, those that involve one of the three discrepant stations being marked by a cross, the remainder by a circle. The effect of the transformation is to move the crosses to the right of the picture and the circles to the left, showing that the plot consists of two distinct clusters.

Fig. 4.5: Semivariogram Plots in D Space



On the strength of these results, it appears that the methodology leads to estimators of spatial covariance which take into account the discrepant behavior of ozone at the three southeasternmost stations. We do not consider any further here the consequences of this for the design of an ozone monitoring network, but one obvious conclusion is that it is necessary to pay special attention to monitoring ozone in this part of the network.

5. CLIMATE EXAMPLE

For our second example, we consider temperature measurements from 138 stations scattered over the continental United States. The stations form part of the Historical Climatological Network (HCN) and have been selected for their long period of continuous high-quality data. As ultimate objectives of the study, we might consider how to measure climate trends taking into account that these are not the same everywhere — it may be necessary to consider a model in which the climate trend varies smoothly with spatial location, but the accurate estimation of such a model would then require that one should take into account spatial correlations as well as spatial variability in means and trends. With this in mind, we now focus exclusively on the spatial correlations and defer to a future paper the investigation of the consequences of this for climate change. A published example in which spatial analysis has been used to inform decisions about climate change is the paper by Handcock and Wallis (1994).

From this data set, a correlation matrix was constructed based on 40 years of annual average temperatures at each of the 138 stations. The 40 years were selected as those for which a reasonably complete record was available at all 138 stations. This correlation matrix will then be treated as a sample correlation matrix on the assumption that observations from different years form independent, identically distributed random vectors. Obviously this appraoch will fail to take into account the effect of both temporal correlations between years, and long-term temperature trends such as might arise from global climate change, but our aim in the present study is to uncover spatial structure in the data rather than to produce a definitive analysis taking into account all aspects of variability.

For this data set, the same kinds of models were fitted by the same methods as for the ozone data. A key issue is again the order $i_1, i_2, ...$, in which the possible centers of the radial basis functions are introduced into the model (cf. (2.9)) and in this case there is even less scope to determine an optimal ordering. Not only are the combinatorial problems of subset selection much greater with 138 stations than with 21, but the time taken to compute each value of the log likelihood (which includes factorizing a 138×138 matrix) is much greater, making the whole procedure extremely computationally intensive. For this reason, a single ordering of the centers was determined prior to any model fitting, mainly chosen so that at each stage of the model fitting process, the centers in the model provide reasonable geographical coverage over the whole region being studied.

Based on this, and employing a log likelihood criterion for selecting the number m of centers included in the model, an initial model selection was made with $m = 21$. [At

the time of writing this does not take account of the cross-validation criterion discussed in sections 3 and 4, and it is hoped to include that in a later version of the paper.] For this data set the Matérn covariance function again provided a reasonable fit, but it was found essential to include a parameter representing the "nugget effect".

Fig. 5.1 shows plots of the sample semivariogram in both G and D space, with the fitted Matérn curve for the latter. Although the transformation from G to D space unquestionably improves the fit as measured by the log likelihood, it must be admitted that there is not much evidence from this in Fig. 5.1, especially when we contrast with this with the very noticable improvement seen with the ozone data between Figs. 4.4 and 4.5! Maps of the G and D space are shown in Fig. 5.2 and it is evident that the main effect of the transformation is to pull a group of stations in the southwestern states (California, Nevada, Arizona) away from the rest of the country.

However there is also evidence in Fig. 5.1 that the semivariogram is *decreasing* at very large distances. As we have noted in Section 2, the Sampson-Guttorp apprach does not allow for the possibility of a non-monotone semivariogram, but our approach via the Bessel representation (2.12) does create this possibility. After some further experimentation a Bessel model with four components was fitted, with results shown in Fig. 5.3. The map of the D space is similar to that in Fig. 5.2, but the semivariogram plot now shows evidence of two distinct clusters of points, with the semivariogram flat or decreasing in the right-hand cluster.

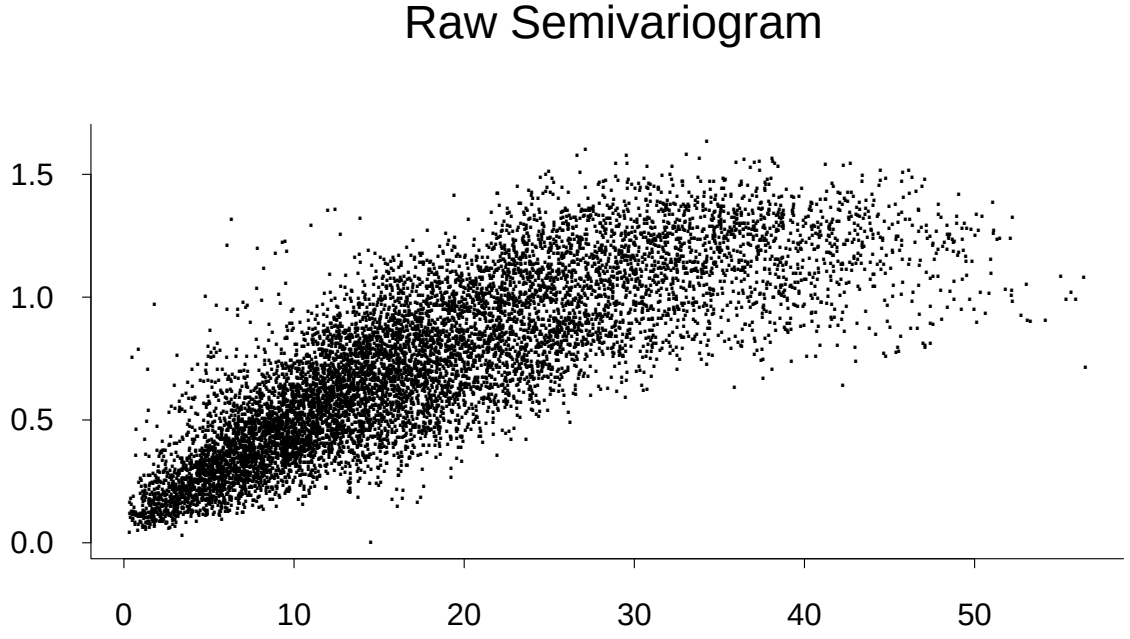
In discussing these results with a climatologist colleague [2] it was suggested that there might be a climatological explanation based on the patterns of air circulation over the continent. There is a tendency for weather patterns to move northwards up the west coast of the USA, then eastwards over the northern Rockies, and then to fan out over the rest of the country. This might well induce a negative correlation between the region soutwest of the Rockies and the rest of the country. However, it was also pointed out that this pattern of air circulation is much more prevalent during the summer months than the winter.

This suggested a re-analysis based on sample correlations computed separately for the summer (June, July, August) and winter (December, January, February) data, with results shown in Figs. 5.4 and 5.5 respectively. The distortion of the map created by the southwestern states is indeed much greater in the summer than the winter, and the evidence for the semivariogram to be decreasing at large lags is also much greater for the summer than for the winter. These results therefore reinforce the climatological explanation.

In fact, examination of raw sample correlations shows a much higher proportion of negative correlations (usually involving the three southwestern states) than could be explained by chance variation on the assumption that the true correlations are always non-negative. This however points to a limitation of the whole approach taken in this paper. If it were

[2] Professor Peter J. Robinson of the Department of Geography, University of North Carolina at Chapel Hill, whose input is gratefully acknowledged

Fig. 5.1: Semivariogram Plots for Climate Data



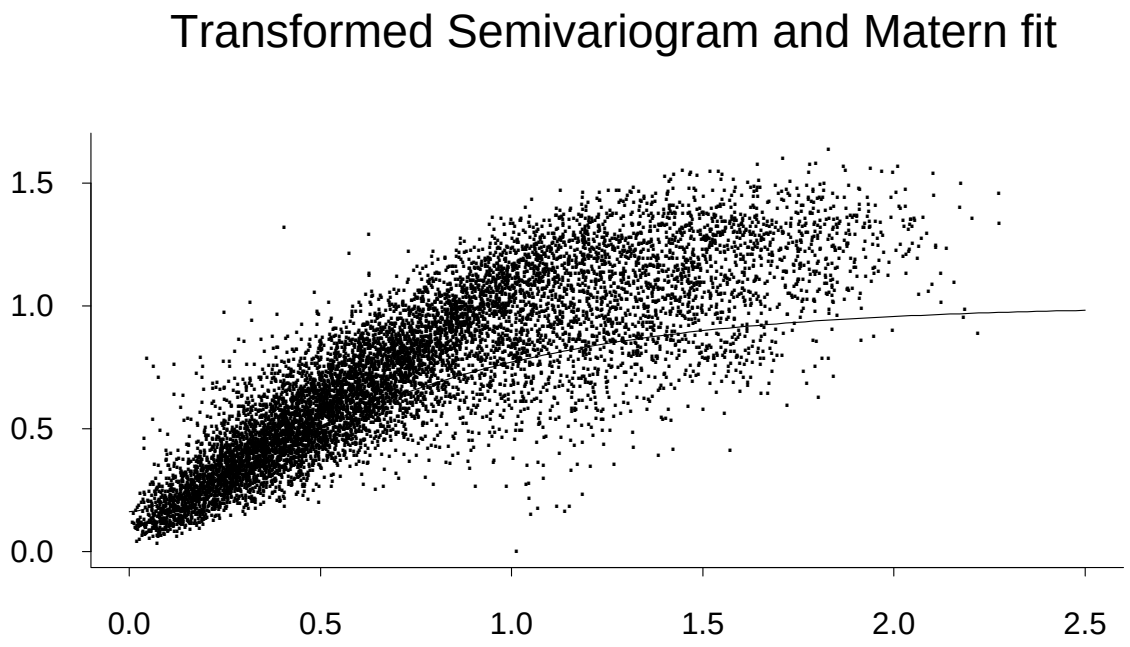
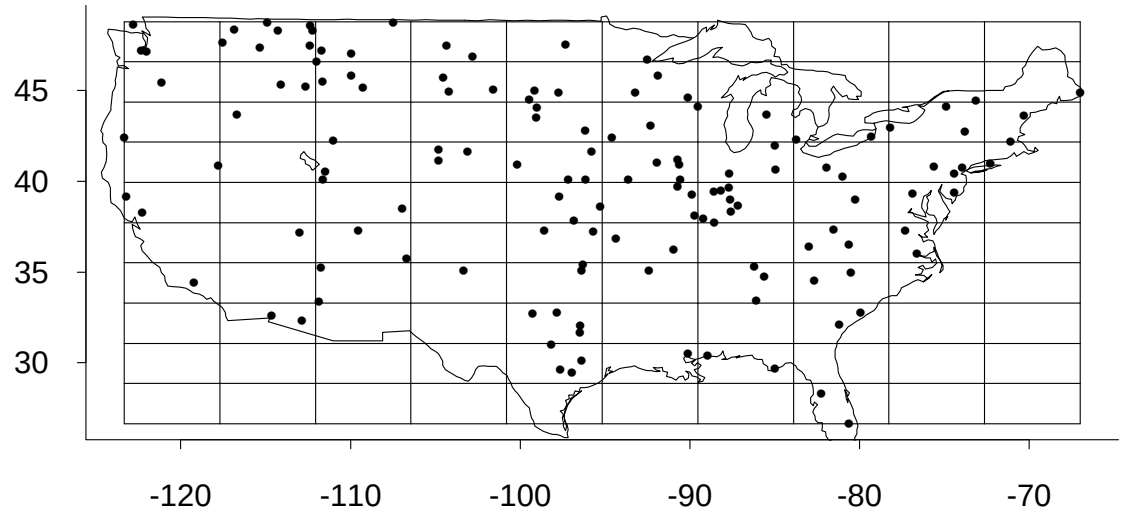


Fig. 5.2: Original and Transformed Maps



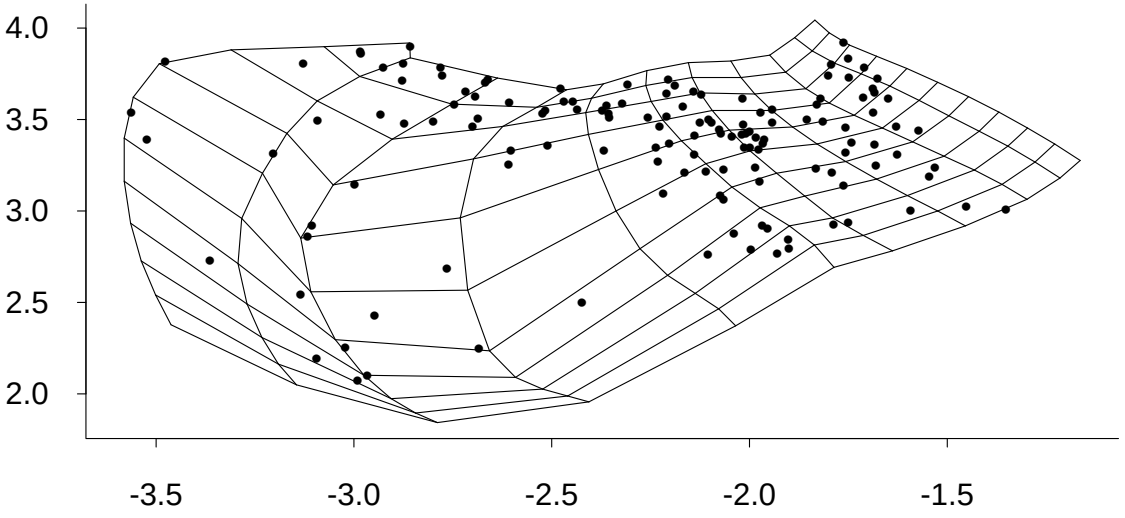
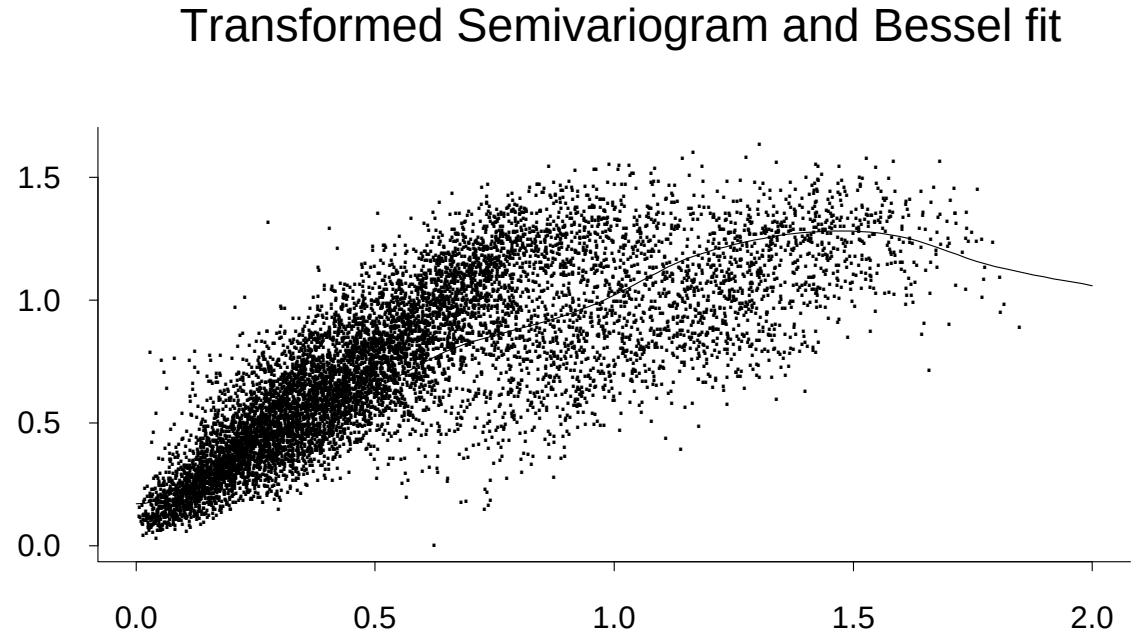


Fig. 5.3: Bessel Semivariogram Fit



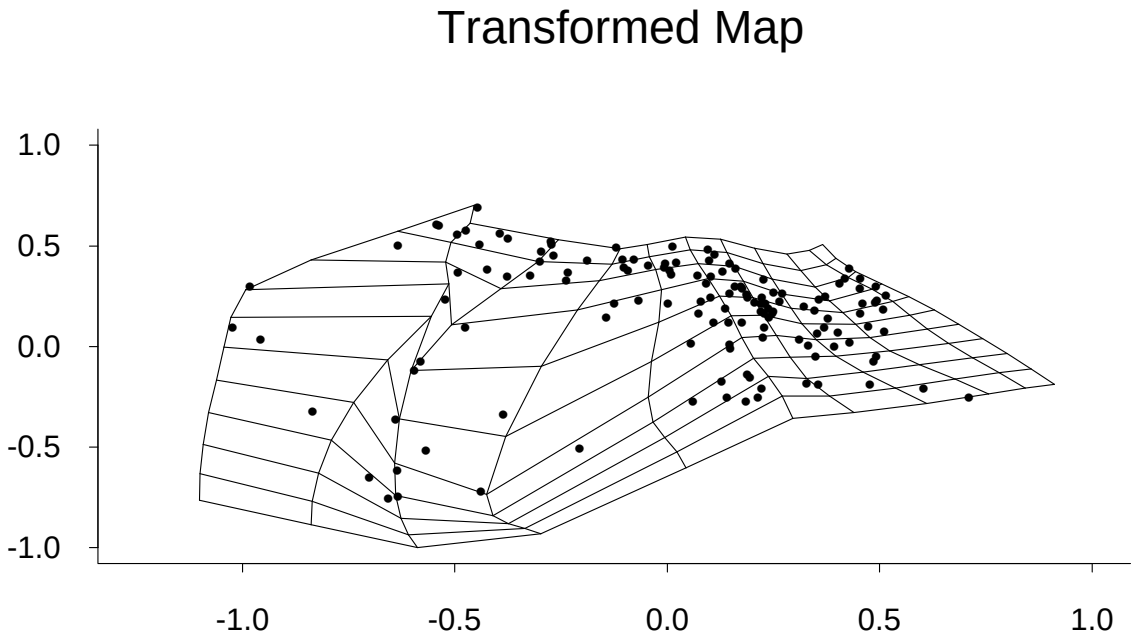
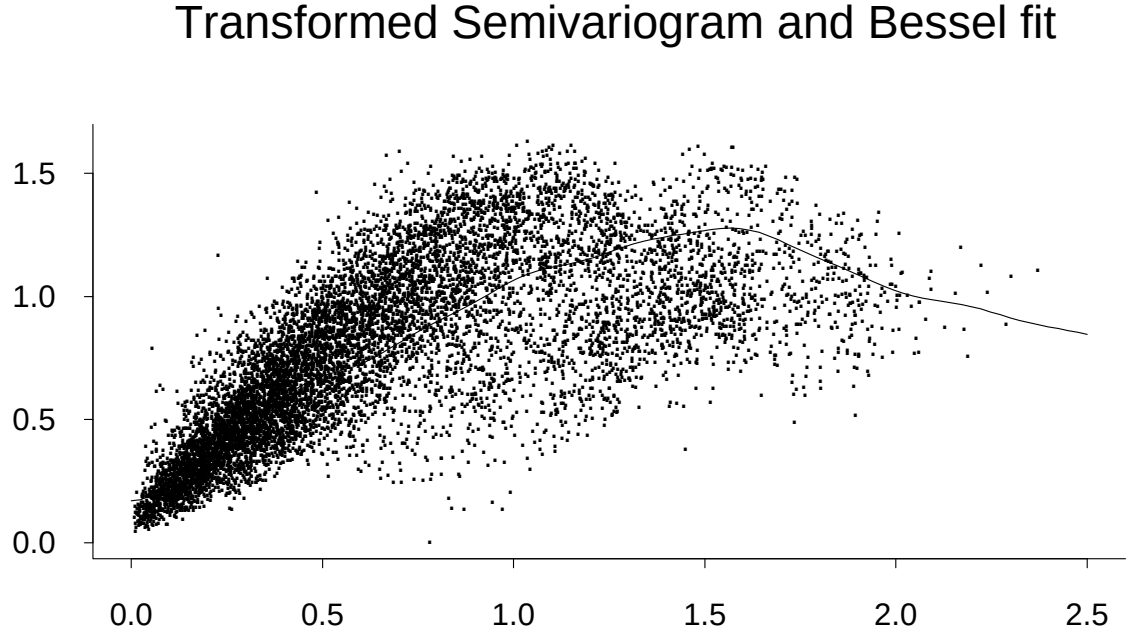


Fig. 5.4: Bessel Fit for Summer Data



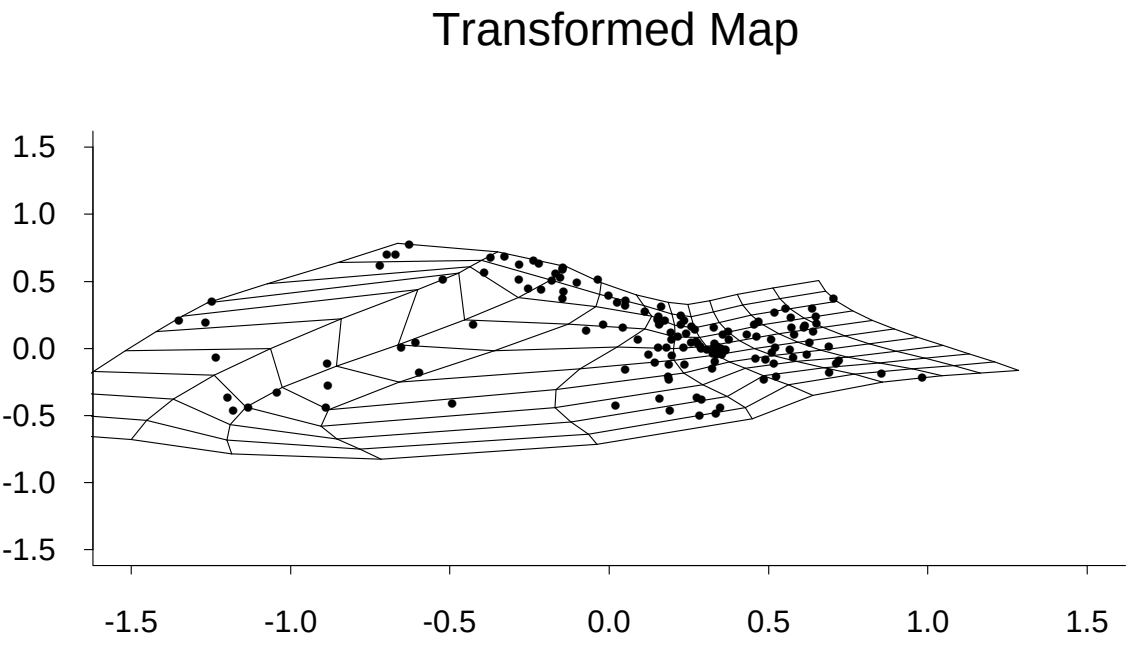
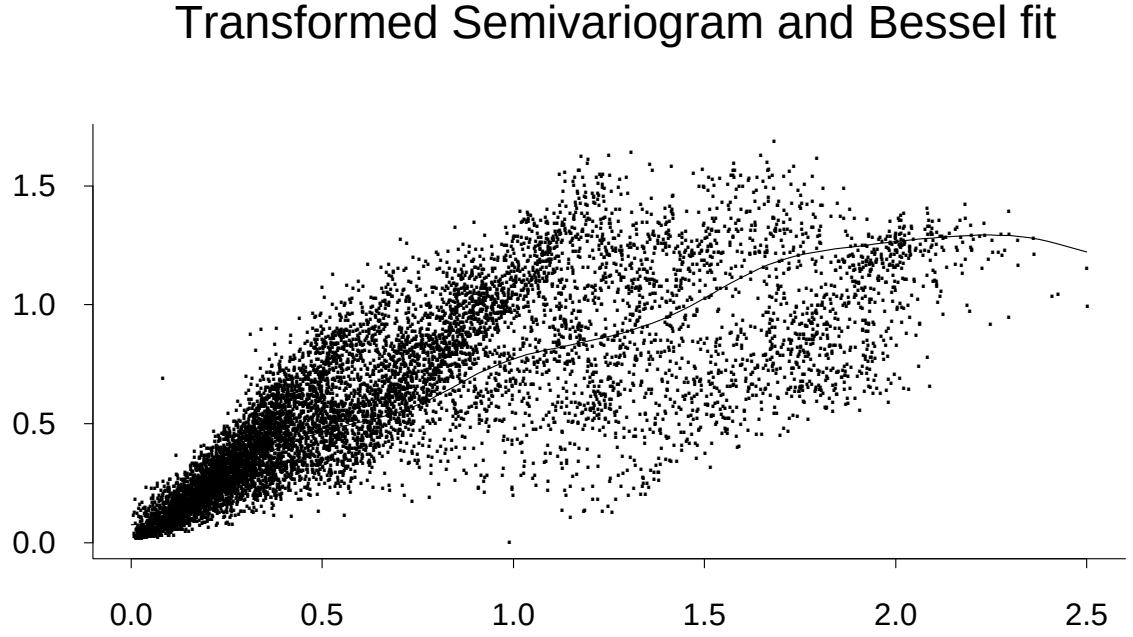
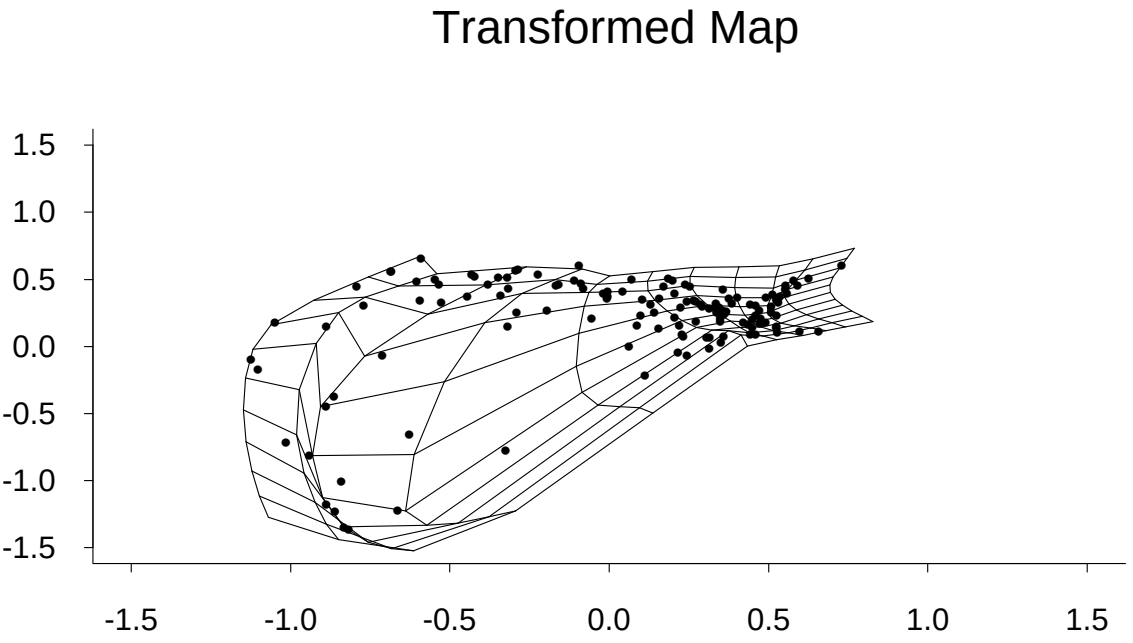


Fig. 5.5: Bessel Fit for Winter Data





indeed the case that in one part of the country, spatial correlations are negative over certain distance ranges, while in another part, correlations are always non-negative regardless of the distance between two stations, then we could not expect the model (1.2) to capture this adequately. To do that, we would need a model which explicitly allowed for the homogeneous semivariogram function γ_0 to be different in different parts of the space.

The other feature of this example that should be made clear is that, for models and data sets of the size being considered here, the question of uniqueness of local maxima of the log likelihood function is something that definitely needs to be considered. Indeed, all the evidence is that the local maxima are not unique, since re-runs of the same model based on different starting values typically result in different estimates of the coefficients of the radial basis functions. Thus in this case the concerns originally raised by Warnes and Ripley (1987) are seen to be valid. In most cases, when two fits of the same model produce different answers, the log likelihood values are very close, and the resulting maps and semivariogram plots also very similar in appearance. Nevertheless, the algorithm sometimes stops at parameter values which are clearly a long way from optimality. Therefore as a practical measure, it is recommended that the same model be re-run from several different starting values before accepting any of the model fits as definitive.

Summarizing, the results for the climatological data are considerably more complicated than those for the ozone data. Nevertheless the fitted models provide considerable insight into the true spatial structure of the data, a statement which is reinforced by the interpretation in terms of streamflow patterns.

6. A NUMERICAL EXPERIMENT

The difficulties in selecting the right sequence of centers for the radial basis function representation, combined with the non-uniqeness of local maximum likelihood estimates, show that it is still highly problematic to identify a single "best model" using the methods we have outlined in this paper. Does it matter? More broadly, do we have evidence that the complicated approaches outlined in this paper really improve on very simple approaches to the estimation of spatial structure, such as an exponentially decaying covariance fitted without any regard for possible nonstationarity or nonisotropy of the data?

A number of variations of the following experiment were tried. Take one of the fits produced by the model, and call it model 1. Now take a competing fit, such as one produced by a slightly different selection of centers or possibly a re-fit under the same model when this does not produce the same result. Call this model 2. Now as a "straw man" alternative, fit an exponentially decaying covariance under the assumption that the true structure is stationary isotropic. This is model 3. If the nonstationary models are working well, and if their nonuniqueness is not to be too much of a concern, then we would like some assurance that in terms of their performance, models 1 and 2 are much closer to each other than either of them is to model 3.

One way to assess the performance of a model is to use it for spatial prediction (kriging). So the following experiment was conducted. Let us assume that model 1 is the true spatial covariance. Construct kriging estimators, at a number of unmonitored sites, using each of the models 1, 2 and 3. Calculate the mean squared prediction errors (MSPEs), for each of the three estimators, when model 1 is true. Of course, the MSPEs associated with model 1 will be the smallest — this is simply a re-expression of the optimality criterion that defines the kriging estimator, i.e. that it minimizes the MSPE when the assumed model is correct. However if things are working well, model 2 should not be a great deal worse than model 1, whereas model 3 might well be very much worse. The mean over all predicted locations of the ratio

$$\frac{\text{MSPE using kriging estimator from model } j}{\text{MSPE using kriging estimator from model 1}}, \tag{6.1}$$

computed separately for $j = 2$ and $j = 3$, may then be taken as an indicator of the relative performance of the two "wrong" models.

The initial results of this analysis have produced rather sobering conclusions. Here is a sample of results, where in each case a particular choice has been made, which will not be elaborated, for models 1 and 2.

We have to agree on a set of unmonitored sites for which predictions will be sought. The reader will observe that in each of the maps drawn in this paper, the spatial locations have been represented within a $10\times$ 10 grid. Therefore one obvious choice for the prediction sites is the set of 81 interior grid points.

With this choice of models and prediction sites, for the ozone data, the ratio (6.1) comes to 1.47 for $j = 2$ and 1.17 for $j = 3$. The same results for the climate data produce 1.21 for $j = 2$ and 1.18 for $j = 3$. In other words, model 2 performs *worse* than model 3, dramatically so in the case of the ozone data.

However on further reflection, perhaps this result should have been expected. In both data sets, there are substantial regions where data sites are either absent or very sparse — over Lake Michigan in the ozone example, and in much of the underpopulated western regions of the USA for the climate example. A model which relies on picking up local fluctuation in the spatial covariances cannot be expected to perform well in regions where there are very few stations, and it is quite conceivable that the exponential isotropic model would provide a better broad-brush representation of the covariances between these regions and the observed monitoring stations.

Another way to assess the predictive performance of the models is to see how well they reproduce the results at the monitoring stations themselves. This has something in common with the cross-validatory approach worked out in section 3 (though for the analysis considered here, it does not involve re-fitting the model). Four separate runs of the kriging experiment were performed, in each case using three quarters of the data points

to predict the remaining one quarter. This was repeated for each of models 1, 2 and 3, and the results evaluated as above, assuming model 1 is the true model. The ratio (6.1) was calculated, averages over all data points and all four runs of the experiment.

For the ozone data, the result was 1.03 under model 2 and 1.11 under model 3. So at least in this case, we have established that model 2 is a better fit than model 3! For the climate data, the respective values were 1.09 and 1.11 — again evidence in favor of model 2, though admittedly not very convincing.

In future studies of this nature it is intended to repeat the experiment using different choices of models 1, 2 and 3, and in particular to integrate the results better with the cross-validation criterion of Section 3 (which has not been used at all in selecting the models for this section). In the meantime, the results serve as a warning against overenthusiastic fitting of very complicated models. Ultimately one would obviously hope that the models do give good predictions over unmonitored regions of the sampling space!

7. SUMMARY AND CONCULSIONS

The analysis of this paper has deliberately focussed on a specific part of the problem, namely how to estimate spatial correlations without taking into account spatial variations in either the mean or the variance, let alone the possibility of different trends at different places (which is obviously an issue with the climate example, and may well be important in ozone monitoring as well). Another issue is the development of spatial-temporal models which make due allowance for temporal as well as spatial correlations. These topics must await future research, but the development of an adequate likelihood-based approach will greatly facilitate such research.

As far as the present paper is concerned, I believe that the results demonstrate the feasibility of a maximum likelihood approach but the difficulties associated with model selection, and those created by the non-uniqueness in many cases of the local maximum of the likelihood function, must not be ignored. We have seen that the cross-validation approach tends to select models of much lower order (i.e. fewer centers in the radial basis function representation) than a likelihood ratio criterion, and the results of section 6 show that the the issue of prediction into unmonitored regions of the sample space must be considered very carefully. These considerations would seem to point towards caution in the fitting of highly nonlinear transformations. Finally our discussion of negative correlations in the climate example points towards a possible limitation of this whole class of models.

REFERENCES

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

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.

Casdagli, M. (1989), Nonlinear prediction of chaotic time series. *Physica D* **35**, 335-356.

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

Green, P.J. and Silverman, B.W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall, London.

Guttorp, P., Sampson, P.D. and Newman, K. (1992), Nonparametric estimation of spatial covariance with application to monitoring newtork evaluation. Chapter 3 of *Statistics in the Environmental and Earth Sciences*, eds. A. Walden and P. Guttorp, Halsted Press, New York.

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

Handcock, M.S. and Wallis, J.R. (1994), An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association* **89**, 368–378.

Healy, M.J.R. (1968), Algorithm AS6: Triangular decomposition of a symmetric matrix. *Applied Statistics* **17**, 195-197.

Journel, A.G. and Huijbregts, C.J. (1978), *Mining Geostatistics.* Academic Press, London.

Judd, K. and Mees, A.I. (1995), On selecting models for nonlinear time series. Technical report, University of Western Australia.

Lewis, T. (1989), Discussion of the paper "Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource", by J. Haslett and A.E. Raftery. *Applied Statistics* **38**, 29.

Loader, C. and Switzer, P. (1993), Spatial covariance estimation for monitoring data. In *Statistics in the Environmental and Earth Sciences*, P. Guttorp and A. Walden (eds.)

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.

Mardia, K.V. and Watkins, A.J. (1989), On multimodality of the likelihood in the spatial linear model. *Biometrika* **76**, 289-295.

Nychka, D. and Royle, A. (1996), Forthcoming technical report on ozone monitoring networks. National Institute of Statistical Sciences, Research Triangle Park, North Carolina.

Oehlert, G.W. (1995), Shrinking a wet deposition network. *Atmospheric Environment*, to appear.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1986), *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press.

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

Smith, L.A. (1992), Identification and prediction of low-dimensional dynamics. *Physica D* **58**, 50-76.

Warnes, J.J. and Ripley, B.D. (1987), Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika* **74**, 640-642.