On space-time covariance for geostatistical data

Urban Hjorth
Department of Mathematical Statistics
Chalmers, Gothenburg

Abstract

A method for extending spatial covariance functions to space and time covariance is studied and its use is examplified with interpolation and prediction from irregular observations to the values in a grid. We consider in particular an inverse linear spatial covariance function and effects of dimension reduction via principal components on the prediction error variances.

1 Introduction

Geostatistical data are multivariate due to the usually large number of observation points and sometimes also due to multivariate observations in each point. Since the measured fields evolve as processes in time, series of such data typically also have time dependency. There is interest in predicting, interpolating and reconstructing such fields and often this is done for a grid of points in order to produce maps. This usually goes under the name of Kriging. The covariance structure, or the semi-variogram (Cressie, 1991), is important for doing this but there is a shortage of tractable covariance structures in space-time, although some recent progress has been made. We need expressions of reasonable simplicity and when possible related to some physical mechanism.

We will discuss covariance models first and then also consider some methods of dimension reduction and the corresponding covariance structure.

2 Some covariance models

We will limit the discussion to stationary and isotropic covariance models. This means that if $X(\mathbf{r}, t)$ is a random field with space coordinate $\mathbf{r}$ and time coordinate t then $Cov(X(\mathbf{r}_1 + \mathbf{r}, t_1 + t), X(\mathbf{r_1}, t_1)) = \mathbf{C}(\mathbf{r}, t) = C(r, t)$, where $r = \sqrt{\mathbf{r}'\mathbf{r}}$. We use bold notation $\mathbf{C}$ for the version defined in $R^d \times T$. When fields are non isotropic we can sometimes eliminate this by different scaling in orthogonal directions, so the models can also cover this case. In some situations it can also be motivated to extend the spatial concept in order to cover covariates such as surface level, water depth, prevailing wind etc when differences in such factors affect the covariances, see Nordgaard-Hjorth (1993). Here we will mainly consider 3-dimensional space.
Whittle (1962) studied diffusion processes driven by noise and taking place in one, two or three spatial dimensions and time. For three dimensions a main conclusion was that this could produce a spatial covariance decreasing as $1/r$ for large r. Based on the equation

$$\frac{\partial X}{\partial t} = -aX + \frac{1}{2}\nabla^2 X + \varepsilon, \tag{1}$$

(containing damping, diffusion and noise) the following spatial covariance types are derived for d spatial dimensions

$d = 1$: $\beta_1 e^{-\beta_2 r}$;

$d = 2$: $\beta_1 K_0(\beta_2 r)$;

$d = 3$: $\frac{\beta_1}{r} e^{-\beta_2 r}$;

where $K_0(x)$ is a Bessel function behaving as $e^{-x}/\sqrt{x}$ for large x and as $-\log x$ for small x. Notice that the variance is infinite for $d = 2, 3$, so the solution is not useful for our purposes without modification, but Whittle provides this by introducing dependency in the noise. When $\beta_2 \to 0$ in three dimensions, corresponding to $\alpha \to 0$ in the differential equation, this produces a $1/r$-behaviour observed in large scale agricultural experiments. This

and some further work by Whittle has inspired later work in the area. See eg Jones and Zhang (1997) where covariance functions are derived from differential equations and expressed as fairly complex integrals.

One class of so called separable covariance models can be constructed as products of spatial covariance functions and temporal covariance functions. The result becomes positive definite and is therefore valid provided the functions we multiply are so. In a forthcoming paper Cressie and Huang (1999+) introduce a technique for producing valid non separable covariance functions in space and time from simpler components. The core of this method is to start with a parameterised set of temporal covariance functions $\rho(\omega, t)$, where $\omega = (\omega_1, \ldots, \omega_d)'$ is the parameter. Then define a positive function $k(\omega)$ with finite integral and compute $\mathbf{C}(\mathbf{r}, t) = \int e^{i\mathbf{r}'\omega}\rho(\omega, t)k(\omega)d\omega$.

This produces positive definite solutions and some closed form expressions are given based on covariance functions and spectral density functions from Matern (1986). Examples in $R^d \times T$, $T =$ time, are

$$C(r,t) = \frac{1}{(t^2+c_0)^{d/2}} exp(-\frac{r^2}{t^2+c_0});$$

$$C(r,t) = \frac{1}{(t^2+c_0)^{d}}(1 - \frac{r^2}{(t^2+c_0)^2})^{-\frac{d+1}{2}};$$

$$C(r,t) = exp(-(t^2+c_0)r^2 - a_0t^2),\ c_0 > 0, a_0 > 0,$$

and scale transformed versions of these. In each case also t^2 replaced by $|t|$ gives a valid covariance function.

3 From space to space-time

As an intermediate more inference based approach we consider a situation where a spatial (stationary and isotropic) covariance function is estimated from simultaneous data or otherwise modelled. Then a differential equation will be used to extend this to a space-time covariance function. In this way the physical motivation behind the differential equation can be combined with spatial evidence from the data. We will apply this strategy with the same differential equation as in Whittle (1962), but without a need to specify the exact nature of the noise part except that we consider it as white in time.

In fact not even this is necessary if we can prove that the resulting covariance is positive definite but it motivates the step from the differential equation for the process to the corresponding differential function for the covariances.

Thus, let a field $X(\mathbf{r},t)$ have the spatial covariance $\mathbf{C}(\mathbf{r},0)$. Let the field be differentiable (as limits in the mean) and follow equation (1) in three dimensions. We allow for scaling by introducing a parameter b in front of the second derivatives and get the differential approximation

$$X(\mathbf{r},t+h) = X(\mathbf{r},t)(1-ah)+\frac{bh}{2k^2}(\sum_1^6 X(\mathbf{r}+\mathbf{k}_i,t)-6X(\mathbf{r},t))+\varepsilon(\mathbf{r},t+h)+o_p(h)$$

where $\mathbf{r}+\mathbf{k}_i$ are the six neighbour points at distance k in the x,y,z directions and the innovation ε is white in time or at least uncorrelated to the field at time 0 which is what we need. Multiplying with $X(\mathbf{0},0)$ and taking expected values and letting $h,k \to 0$ we arrive at the corresponding differential equation for the covariances

$$\frac{\partial \mathbf{C}}{\partial t} = -a\mathbf{C} + \frac{b}{2}(\frac{\partial^2 \mathbf{C}}{\partial x^2} + \frac{\partial^2 \mathbf{C}}{\partial y^2} + \frac{\partial^2 \mathbf{C}}{\partial z^2}). \tag{2}$$

This equation has the solution

$$\mathbf{C}(\mathbf{r},t) = e^{-at}\int\int\int \mathbf{C}(\mathbf{u},0)\frac{1}{(2\pi bt)^{3/2}}e^{-\frac{((\mathbf{u}-\mathbf{r})'(\mathbf{u}-\mathbf{r})}{2bt}}d\mathbf{u},\ t>0, \tag{3}$$

where the kernel used is the probability density of a $N_3(\mathbf{0}, bt\mathbf{I})$-distributed variable. Usually polar coordinates will make computations more easy. For a negative t we use $|t|$ instead. Since we have not verified that every spatial covariance function can arise from this kind of diffusion mechanism we state the following:

Theorem: If $\mathbf{C}(\mathbf{r},0)$ is positive definite in space then $\mathbf{C}(\mathbf{r},t)$ given by (3) is positive definite in space and time.

Proof: We show that the solution (3) has a non-negative spectral measure. For t fixed, the solution is a convolution between $\mathbf{C}(\mathbf{u},0)$ and a Gaussian function and their spectra are therefore multiplied. The spectral measure of $\mathbf{C}(\mathbf{u},0)$ is time independent and non negative. By direct integration with respect to x,y,z,t the function $e^{-a|t|}|t|^{-3/2}exp(-(x^2+y^2+z^2)/2b|t|)$ has a non negative Fourier transform.

This proves that the solution is always valid. We can also use that linear combinations with positive weights of covariance models are also valid and in particular extra variance can be added to $\mathbf{C}(\mathbf{0}, 0)$, the so called nugget effect used to describe observational error and very local variance which is uncorrelated at the time and space distances used in observations. We are of course also allowed to use any covariance model from d-dimensional space as a model in lower dimension by always setting the extra coordinates as zero. We give some examples of basic functions.

3.1 Gaussian space-covariance

Let the spatial covariance at time lag zero be $C(r, 0) = e^{-dr^2}$. Then the space-time solution (3) is easy to integrate and becomes

$$C(r, t) = e^{-at} \frac{e^{-\frac{dr^2}{2bdt+1}}}{(2bdt + 1)^{3/2}}, \ t > 0.$$

More generally we can allow for variance and write a four parameter model as

$$C(r, t) = \frac{\sigma^2}{(\beta t + 1)^{3/2}} e^{-at - \frac{\gamma r^2}{\beta t+1}}, \ t > 0. \tag{4}$$

3.2 Negative exponential space-covariance

If $C(r, 0) = e^{-\alpha r}$ we get a somewhat more involved result. A solution with one-dimensional space can be found in Jones and Zhang, (1997). In 3 dimensions we get for $t > 0$

$$\mathbf{C}(\mathbf{r}_0, t) = e^{-at} \int\int\int e^{-\alpha r} \frac{1}{(2\pi bt)^{3/2}} e^{-\frac{(\mathbf{r}-\mathbf{r}_0)'(\mathbf{r}-\mathbf{r}_0)}{2bt}} d\mathbf{r}.$$

Using spherical symmetry we may choose $\mathbf{r}_0 = (0, 0, r_0)'$ and with polar coordinates we get

$$C(r_0, t) = \frac{e^{-at}}{(2\pi bt)^{3/2}} \int\int\int e^{-\alpha r - \frac{r^2 - 2r_0 r \, sin\theta + r_0^2}{2bt}} r^2 cos\theta \, d\phi d\theta dr.$$

This gives for $t > 0$

$$e^{at} C(r, t) = \sqrt{\frac{bt}{2\pi}} e^{-\frac{r^2}{bt} - \frac{\alpha^2 bt}{2}} \frac{e^{\alpha r} - e^{-\alpha r}}{r} + \Phi(\frac{r - \alpha bt}{\sqrt{bt}}) + \Phi(\frac{-r - \alpha bt}{\sqrt{bt}})$$

$$-\frac{\alpha bt}{r}\left(\Phi(\frac{r-\alpha bt}{\sqrt{bt}})-\Phi(\frac{-r-\alpha bt}{\sqrt{bt}})\right), \tag{5}$$

where $\Phi(x)$ is the distribution function of the standardised normal distribution. We can complete this with a variance factor σ^2.

3.3 Inverse linear case

The long range behaviour of type $1/r$ was studied by Whittle and has some special interest since it was observed in real field data. Perhaps the simplest type of function allowing this is $1/(1+dr)$, but to use this we must know that it is positive definite.

Proposition The function $1/(1+r)$ is positive definite in R^d.

Proof: Consider the Laplace transform (Stand. Math. 1974, p. 510).

$$\frac{1}{\sqrt{s}+1}=\int_0^\infty e^{-s\lambda}(\frac{1}{\sqrt{\pi\lambda}}-e^{\lambda}\frac{2}{\sqrt{\pi}}\int_{\sqrt{\lambda}}^\infty e^{-u^2}du)d\lambda=\int_0^\infty e^{-s\lambda}dF(\lambda).$$

The measure $dF(\lambda)$ is non-negative since (extracting e^{λ} and substituting x^2 for λ)

$$h(x)=\frac{1}{x}e^{-x^2}-2\int_x^\infty e^{-u^2}du\to 0,\ x\to\infty,$$

and

$$h'(x)=-\frac{1}{x^2}e^{-x^2}<0.$$

From the first expression follows that

$$\frac{1}{1+||\mathbf{r}_i-\mathbf{r}_j||}=\int_0^\infty e^{-\lambda||\mathbf{r}_i-\mathbf{r}_j||^2}dF(\lambda)$$

and

$$\sum_i\sum_j a_i\bar{a}_j\frac{1}{1+||\mathbf{r}_i-\mathbf{r}_j||}=\int_0^\infty\sum_i\sum_j e^{-||\mathbf{r}_i-\mathbf{r}_j||^2\lambda}a_i\bar{a}_j dF(\lambda)$$

is positive since $e^{-\lambda\tau^2}$ is a positive definite function.
Notice that $1/(1+r)$ has a spectral measure but no Fourier transform.

Again, using polar coordinates in (3) and integrating over ϕ and θ, we get the following expression for the space-time covariances when $C(r,0)=1/(1+dr)$.

$$C(r,t) = e^{-at} \int_0^\infty \frac{u}{1 + d\sqrt{bt}u} \frac{\phi(u - u_0) - \phi(u + u_0)}{u_0} du, \; t > 0, \qquad (6)$$

where $u_0 = r/\sqrt{bt}$ and we have substituted the corresponding transformation for u. Here ϕ is the standard normal frequency function, $\phi(x) = exp(-x^2/2)/\sqrt{2\pi}$.

When $r \to 0$ also $u_0 \to 0$ and $(\phi(u - u_0) - \phi(u + u_0))/u_0) \to 2u\phi(u)$. This gives the time covariance

$$C(0,t) = e^{-at} \int_0^\infty \frac{u}{1 + d\sqrt{bt}u} 2u\phi(u) du. \qquad (7)$$

The last expression can be interpreted as an expected value

$$e^{-at} E[\frac{U^2}{1 + d\sqrt{btU^2}}],$$

where $U \in N(0,1)$.

4 Reduction of dimension in Kriging

Suppose we want to predict/interpolate from an irregular (in space or space-time) set of data (X) to values (Y) in a regular grid at a certain time. We treat X and Y as column vectors and each component is regarded as $X_i = X(\mathbf{r}_i, t_i)$ and $Y_j = X(\mathbf{u}_j, t_0)$ for some random field $X(\mathbf{r}, t)$. However, in this section we first give some general results. For simplicity we assume that expected values are removed unless they are explicitly written.

4.1 Effect of principal components

If X or Y have high dimension we may consider reducing this by using an appropriate number of terms in some series expansion. An obvious candidate is principal components (PC) but other regression based possibilities also exist, in particular when there are natural covariates present related to the spatial position. Let PC(X) and PC(Y) be short for development of observations and grid variables into PC. We have the following alternatives to predict Y

depending on if we use one or the other or both of the developments into principal components:

1.	$X \to Y$	$\hat{Y}_1 = \hat{Y}\|X$
2.	$X \to \mathrm{PC}(Y) \to Y$	$\hat{Y}_2 = \hat{Y}\|XPY$
3.	$X \to \mathrm{PC}(X) \to Y$	$\hat{Y}_3 = \hat{Y}\|PX$
4.	$X \to \mathrm{PC}(X) \to \mathrm{PC}(Y) \to Y$	$\hat{Y}_4 = \hat{Y}\|PXPY$

The last code for predictors is hopefully self explaining. We will see what the four alternatives implies for the prediction error variances and we analyse this under the assumption that the covariance properties are known. Predictions are the optimal linear ones without side condition on the row sums of coefficients. Such conditions are otherwise common in kriging to account for a constant but unknown expected value. If the field is Gaussian it is well-known that the linear predictors are optimal also in the class of all (linear and non-linear) predictors based on the given information. Graphical illustrations are given later in an example.

Some notation for the principal components is given here. Let C_{XX} be the covariance matrix for the observations and C_{YY} the covariance matrix for the grid point variables. For later reference we also introduce the cross covariances between those variable vectors as C_{XY} and $C_{YX} = C'_{XY}$. Let $\lambda_1 > \lambda_2 > \ldots$ be the eigenvalues of C_{XX}, here supposed to be distinct. The corresponding normalised eigenvalues are denoted $g_1, g_2, \ldots$ and $G_r = (g_1, \ldots, g_r)$ is the matrix of the first r of them. In the same way $l_1 > l_2, \ldots$ are the eigenvalues of C_{YY} and $H_k = (h_1, h_2, \ldots, h_k)$ the matrix of the first k eigenvectors for the grid. We have

$$G'_r G_r = I_r,\ G'_r C_{XX} G_r = G'_r G_r \Lambda_r = \Lambda_r$$

where $\Lambda_r = diag(\lambda_1, \lambda_2, \ldots, \lambda_r)$, and accordingly

$$H'_k H_k = I_k,\ H'_k C_{YY} H_k = H'_k H_k L_k = L_k$$

where $L_k = diag(l_1, l_2, \ldots, l_k)$. The optimal restoration of Y from the vector $V_k = H'_k Y$ of the k first PC is given by $H_k V = H_k H'_k Y$.

4.1.1 Directly $X \to Y$

Nothing can beat the optimal direct prediction from observations X to Y when the model is known. Passing over PC or some other expansion with

reduced dimension can only restrict our class of predictors, but the reduction of dimension may have other advantages in terms of computer storage, ease of interpretation etc. Let

$$C = \begin{pmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{pmatrix}$$

be the covariance of $(X'\ Y')'$. The optimal prediction is then

$$\hat{Y}_1 = \hat{Y}|X = C_{YX}C_{XX}^{-1}X, \tag{8}$$

and the corresponding prediction error has covariance matrix

$$C_{Y-\hat{Y}|X} = C_{YY} - C_{YX}C_{XX}^{-1}C_{XY}. \tag{9}$$

See for example Mardia-Kent-Bibby (1997), Chapter 3. Of course the same formula gives prediction of $Y - \mu_Y$ from $X - \mu_X$ in the general case when μ is known and not necessarily zero.

4.1.2 Passing PC(Y)

If we reduce the dimension of Y to k principal components then they are given by $\mathrm{PC}(Y) = H_k'Y$. The prediction of $H_k'Y$ directly from X becomes $H_k'\hat{Y}_1 = H_k'C_{YX}C_{XX}^{-1}X$, where we use the result (8), since a linear function of an optimal prediction is optimal for the same function of the variables (due to projection onto the observation space). From this we use another linear transform to predict Y as

$$\hat{Y}_2 = \hat{Y}|XPY = H_kH_k'C_{YX}C_{XX}^{-1}X = A_2X, \tag{10}$$

with prediction error variance

$$\begin{aligned} C_{Y-\hat{Y}|XPY} &= E(Y - A_2X)(Y - A_2X)'] \\ &= C_{YY} - C_{YX}A_2' - A_2C_{XY} + A_2C_{XX}A_2' \\ = C_{YY} - C_{YX}C_{XX}^{-1}C_{XY} &+ (I - H_kH_k')C_{YX}C_{XX}^{-1}C_{XY}(I - H_kH_k'), \end{aligned} \tag{11}$$

where the last term gives the loss due to excluded components in $\mathrm{PC}(Y)$.

4.1.3 Passing PC(X)

We now look at the route $X \to \mathrm{PC}(X) \to Y$. Let $U_r = G_r'X$ be the first r principal components. The covariance matrix of $(U_r', Y')'$ becomes

$$C_{U;Y} = \begin{pmatrix} \Lambda_r & G_r'C_{XY} \\ C_{YX}G_r & C_{YY} \end{pmatrix};$$

and in analogy with (8) this gives

$$\hat{Y}_3 = \hat{Y}|PX = C_{YX}G_r\Lambda_r^{-1}U_r = C_{YX}G_r\Lambda_r^{-1}G_r'X = A_3X. \tag{12}$$

The corresponding covariance matrix for prediction errors follows from (9) as

$$\begin{aligned} C_{Y-\hat{Y}|PX} &= C_{YY} - C_{YX}G_r\Lambda_r^{-1}G_r'C_{XY} \\ &= C_{YY} - C_{YX}C_{XX}^{-1}C_{XY} + C_{YX}[C_{XX}^{-1} - G_r\Lambda_r^{-1}G_r']C_{XY}, \end{aligned} \tag{13}$$

where the last term gives the extra variance due to passing PC(X).

4.1.4 Passing both PC(X) and PC(Y)

Finally the use of PC for both X and Y means that if $V_k = H_k'Y$ are the first k PC of Y, then $\hat{V}_k = H_k'\hat{Y}$ with $\hat{Y} = \hat{Y}|PX$, and the prediction of Y then becomes $\hat{Y}|PXPY = H_k\hat{V}_k$. Thus from (12)

$$\hat{Y}|PXPY = H_kH_k'C_{YX}G_r\Lambda_r^{-1}G_r'X = A_4X, \tag{14}$$

and the corresponding covariance matrix becomes

$$\begin{aligned} C_{Y-\hat{Y}|PXPY} &= C_{YY} - A_4C_{XY} - C_{YX}A_4' + A_4C_{XX}A_4' \\ &= C_{YY} - H_kH_k'C_{YX}G_r\Lambda_r^{-1}G_r'C_{XY} - C_{YX}G_r\Lambda_r^{-1}G_r'C_{XY}H_kH_k' \\ &\quad + H_kH_k'C_{YX}G_r\Lambda_r^{-1}G_r'C_{XY}H_kH_k' \\ &= C_{YY} - C_{YX}G_r\Lambda_r^{-1}G_r'C_{XY} + (I - H_kH_k')C_{YX}G_r\Lambda_r^{-1}G_r'C_{XY}(I - H_kH_k'), \end{aligned} \tag{15}$$

where the last term is the additional variance of passing both PC compared to (13) where we are passing only PC(X).

4.1.5 Noise propagation

The nugget effect is a name for measurement noise plus very local variation acting as independent. If the covariance model already accounts for this variance, the predictions are already balancing this against the "signal" in an optimal way, but it may nevertheless be of interest to to see how this error propagates to the forecasts. If a covariance model is used which neglects this variance, it can be seen as an extra variance added afterwards. Since all our forecasts can be written as $\hat{Y} = AX$, a nugget with the same variance σ_n^2 everywhere will be responsible for the part $\sigma_n^2 AA'$ of the covariance matrix for the prediction vector if it is already included in the model and otherwise it will add this much to it. (The effect on the prediction errors will be the above variance plus σ_n^2, since both the prediction and the observed grid value will have the nugget and they are independent when we predict to new points. We only describe the variance in the predictions here.) This gives

$$\sigma_n^2 AA' = \sigma_n^2 \begin{cases} C_{YX} C_{XX}^{-2} C_{XY}, \\ H_k H_k' C_{YX} C_{XX}^{-2} C_{XY} H_k H_k', \\ C_{YX} G_r \Lambda_r^{-2} G_r' C_{XY}, \\ H_k H_k' C_{YX} G_r \Lambda_r^{-2} G_r' C_{XY} H_k H_k', \end{cases} \tag{16}$$

in the four cases $X \to Y$, $X \to \mathrm{PC}(Y) \to Y$, $X \to \mathrm{PC}(X) \to Y$, $X \to \mathrm{PC}(X) \to \mathrm{PC}(Y) \to Y$. The notation A^{-2} denotes the inverse matrix times itself. See the example below for a numerical illustration.

4.2 EOF before and after modelling expected values

In geostatistics, theoretical covariance matrices are often replaced by cross products of data. Principal components are then developed, usually under the name of empirically orthogonal functions, EOF. Sometimes series over different seasons are put together and only a general mean is subtracted from them. Sometimes a more detailed climatology is subtracted and the series then represent the anomaly only. The anomaly case gives cross products that are close to statistical covariance functions but when the climatology is not subtracted the effect of a common seasonal variation acting on all the observations in approximately the same way is like adding a large constant to all the elements of a covariance matrix. The effect of this is that the prediction equations will be forced towards weights adding to 1 (if the predicted variables are of the same type as the data and have the same climate

variations or seasonal effects). Thus in this case we approach automatically the Kriging with this side condition on the coefficients. As always, the least squares adapts to the situation covered by the data and with strong seasonal variations it must come close to this variation also in the predicted variable. Compared to a perfect model where we subtract expected values, the effect of forcing a side condition on the predictions is of course an increased average squared prediction error variance. We illustrate the effect on a very small 5×5 covariance matrix $c_{ij} = 1/(1 + |i - j|)$ which can be thought of as covariances for points along a line with a unit distance between. The first 3 (called X) are used to predict the last two (Y) Writing $\hat{Y} = AX$ we get from (8)

$$A = \begin{pmatrix} 0.0625 & 0.0833 & 0.4375 \\ 0.0700 & 0.0800 & 0.2700 \end{pmatrix},$$

and the row sums are 0.5833 and 0.4200. The same computation where a constant d is added to all the elements of C gives the following results for the row sums at a set of different d-values.

d	0	1	5	20	100
row 1:	.58	.84	.95	.99	.997
row 2:	.42	.78	.94	.98	.996

4.3 More interpretable expansion

The first principal components have often some useful interpretation based on where the loadings (components of eigenvectors) are large. For a temperature field in Sweden the north-south differences will show up and also sea-land and altitude will be evident in the largest components if the area is large enough. Instead of trying to find such interpretations of the principal components, it is attractive to try a direct modelling in terms of such effects. Sometimes also time lag and information about advection should be considered in the models which may motivate a splitting into different conditional models but we will not go further into that here. Let $\mathbf{a}_i'$ be a vector of covariates for the observation point (such as geographical position, type of environment etc.) and let the matrix $\mathbf{A}$ have such rows. Let the vector $X(t)$ with elements $X(\mathbf{r}_i, t)$ be written as

$$X(t) = A\beta(t) + Z(t),$$

where $\beta(t)$ contains the time variation of the part which can be explained by the covariates in A just like the PC contains the part which can be explained by components up to the truncation point. Of course, with the same number of components the PC will always cover more of the observed variance in observation points, since this is the optimality property of PC, but instead of difficulties in understanding higher components, the understanding is built directly into the construction of the covariates. The remaining vector $Z(t)$ is regarded as noise or variation which can not be explained by the covariates. If a pattern can be seen in Z, then further appropriate covariates can be searched. Although this modelling in terms of covariates may be preferable we can still have much use of PC development in order to find the strong factors and the corresponding covariates.

5 Example

We illustrate our results by an artificial use of one of the covariance models over an area sized as south Sweden. Observation points are approximately positioned as a set of meteorological stations used some years ago in another study, and the covariance model used is the inverse linear one with reasonable spatial covariances for an autumn temperature field according to estimates made in the eighties. The remaining parameters defined by the differential equation are less well known, and not estimated since due to atmospheric movements this would need a combination of both a model for and data on the wind fields, or at least the pressure gradients defining the geostrophic wind, and this would bring in too much for the present purposes. Since our aim is to illustrate covariance models and the effects of PC, without a particular meteorological application in mind, we will set these parameters so that diffusion (governed by b) and exponential decay with time (governed by a) both have some effect on the pure time correlation. Using 1 km as distance unit we put $d = 1/450$. The time unit is more arbitrary and we have chosen the parameter values $a = 0.1$ and $b = 100$ as an illustration.

The covariance function is shown for positive space and time distance in Figure 1. The pure space covariance is $1/(1 + d\dot{r})$ and follows the frontal borderline from right to left and the pure time covariance given by (7) goes inwards along the right border.

In Figure 2 we show the observation points and the grid on a map of south Sweden. The grid has 19 times 19 grid points and is just about covering the

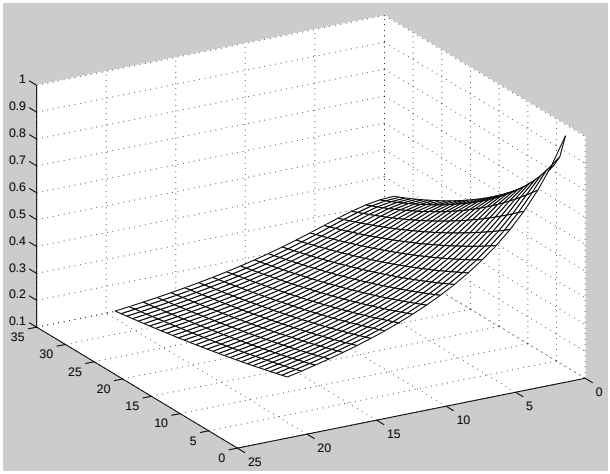

Figure 1: The inverse linear covariance function with parameter setting above. Time difference 0 to 30 times 0.2 inwards, space distance 0 to 20 times 50 km from right to left.

observation points.

Most of the data are simultaneous, but we allow a few of them to be lagged one or two time units as shown in Table 1.
At this point we should nor stress the temperature example too hard because we do not consider the moving atmosphere. Instead we illustrate the capacity of the covariance model to allow time differences due to diffusion mechanisms. We will also demonstrate this by changing from simultaneous to predicted grid values in the illustrations below. (The transport can to some extent be handled by a coordinate system moving with the wind for scales where a common wind can be defined. A time difference will then lead to effects on both time and space differences. In more general situations time-space covariances can be useful as elements in more complicated models of transport for example trajectory based modelling.) Using a Matlab routine we compute the eigenvalues and eigenvectors of the observations and also for the grid values.

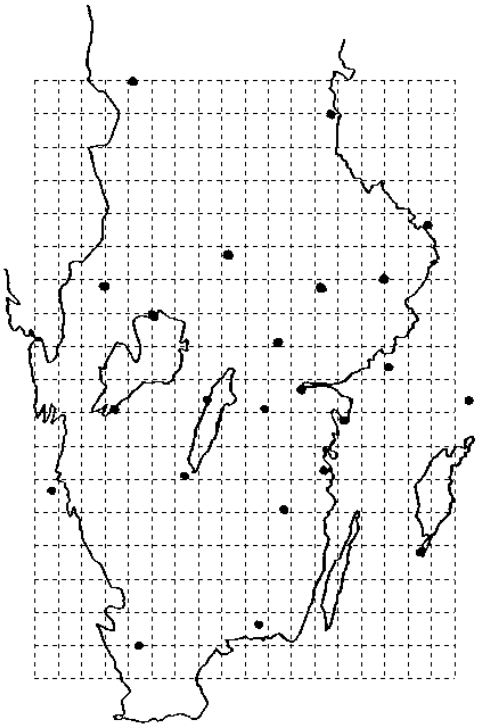

Figure 2: Map of south Sweden with 23 observation points and a grid.

Table 1: Coordinates from a central reference point and time lags. Approximately 5 km distance unit.

x	y	lag	x	y	lag
-24	64	0	-02	-30	0
33	55	0	14	-33	1
62	23	0	-29	-34	0
04	13	2	37	-36	0
49	06	1	32	-51	0
31	03	0	-09	-54	2
-32	03	0	-48	-58	0
-18	-5	0	20	-63	0
18	-13	1	60	-75	0
51	-20	1	13	-92	0
25	-27	0	-22	-104	1
74	-30	0			

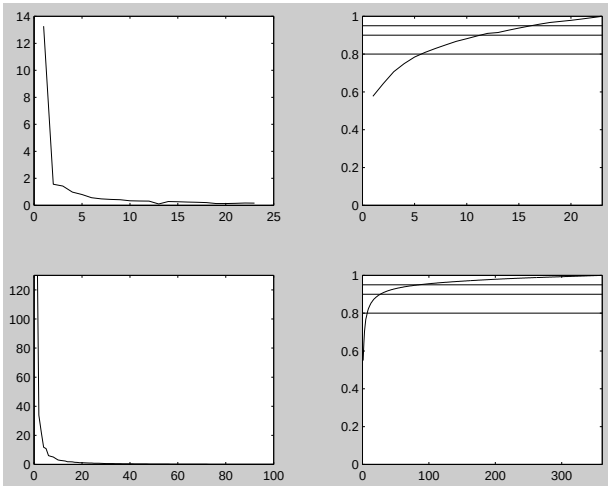

Figure 3: Eigenvalues and their cumulative sums divided by total variance for observations, above, and for grid values, below. Cut off points for 80, 90, 95 % indicated by lines.

In Figure 3 we show the eigenvalues for the set of observations and also the first 100 eigenvalues out of 361 for the set of grid variables. We also plot the cumulative sums and the cut off points covering 80, 90 or 95% of the total variance, but we only give results for the first two cut off points.

Next, the prediction error covariance matrices are computed for direct prediction and prediction via the different PC developments. The variances

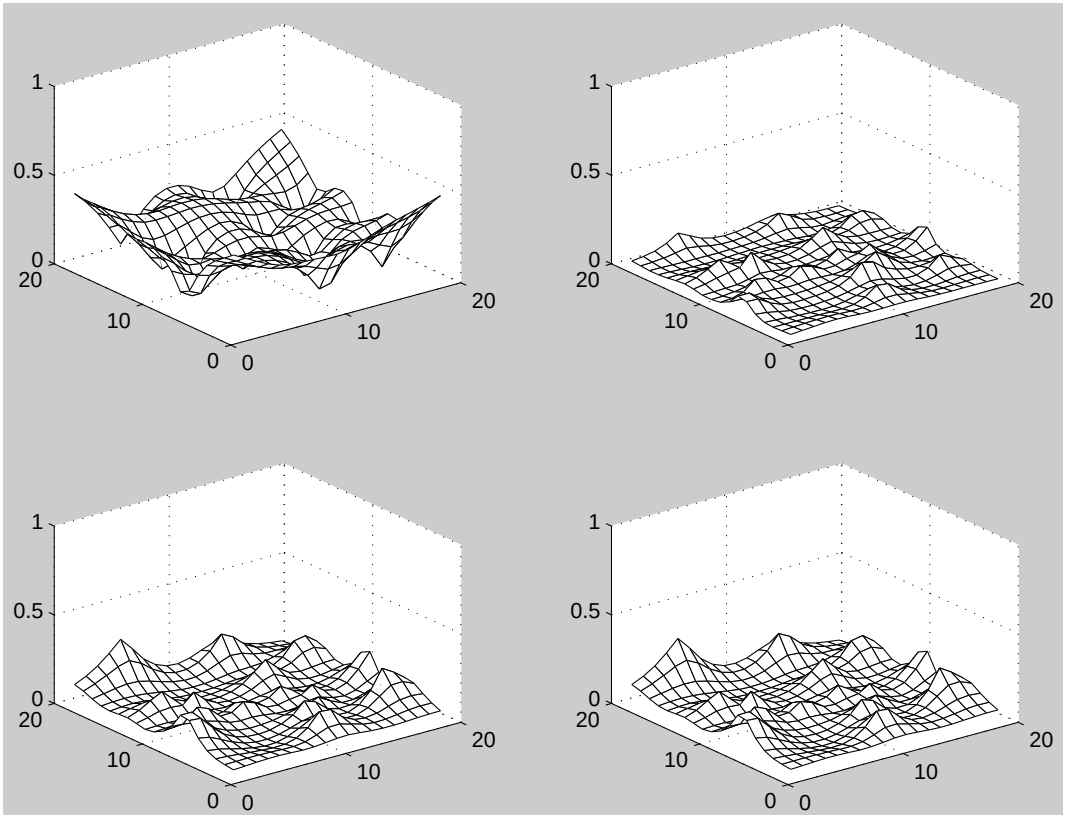

Figure 4: Variance of direct prediction and added variance due to PC(Y), upper right, PC(X), lower left, and both combined, lower right. PC cut off point 80%.

of this covariance matrix are extracted from the diagonal and reshaped into a matrix according to the grid-points they belong to. In Figure 4 the prediction error variance of the optimal direct prediction from observations to the grid, and the added variance for the routes over one or two PC developments using 80% cut of, is displayed for a case with simultaneous observations and grid point values except for a few lagged data. The results show how the optimal prediction of course gives a very small error variance close to the observation points where simultaneous data are available. PC development for the grid variables adds some variance but the big addition comes with the PC for the observations, and it is high around the observation stations but not so high in between. In Figure 5 the same is shown for the 90% cutoff and the structure is similar although the variance additions are smaller.

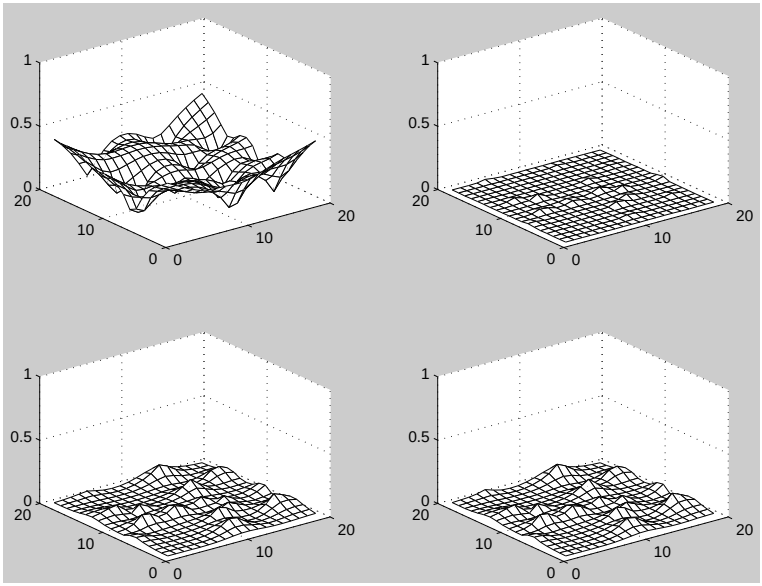

Figure 5: Variance of direct prediction and added variance due to PC(Y), upper right, PC(X), lower left, and both combined, lower right. PC cut off point 90%.

In Figure 6 we stick to the 90% cut off of PC developments and look at prediction 0.2 time units ahead where the result is much more flat. Predicting further ahead, we expect again less differences in the variances and this is verified in Figure 7 where prediction 1 time unit ahead is studied.

Finally we illustrate the variance due to the nugget effect. If we keep the

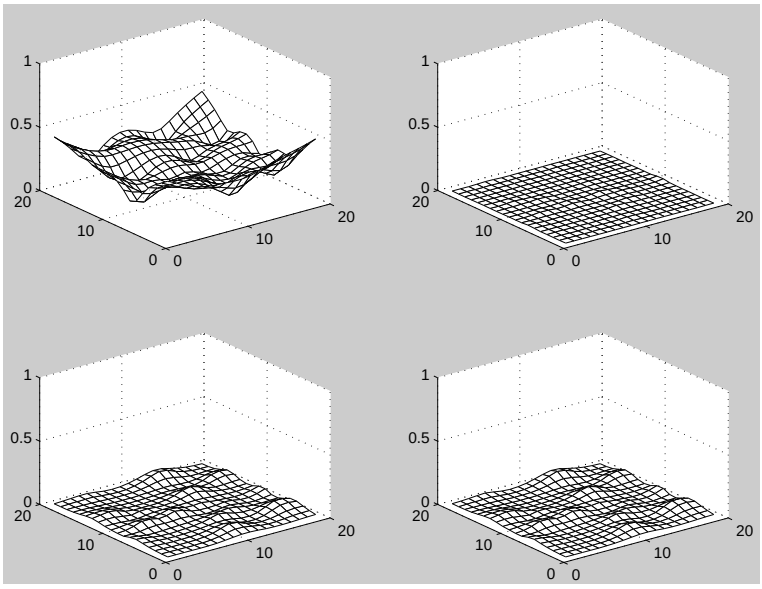

Figure 6: Variance of direct prediction 0.2 time steps ahead and added variance due to PC(Y), upper right, PC(X), lower left, and both combined, lower right. PC cut off point 90%.

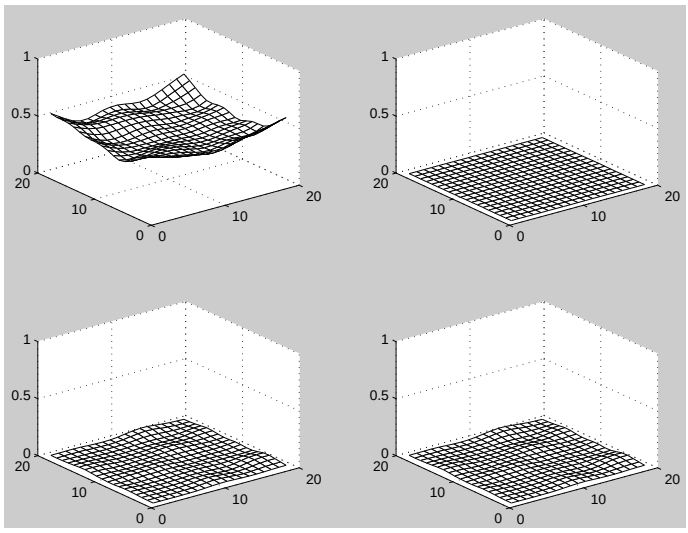

Figure 7: Variance of direct prediction one time step ahead and added variance due to PC(Y), upper right, PC(X), lower left, and both combined, lower right. PC cut off point 90%.

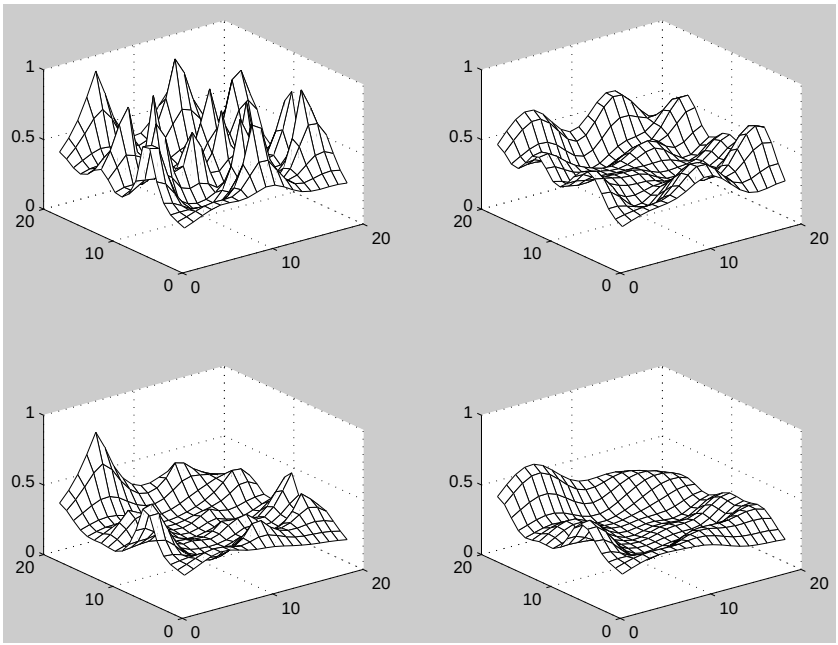

Figure 8: Variance in grid predictions from a thought independent noise in every observation point. Direct pred., upper left, PC(Y) upper right, PC(X) lower left, and both combined, lower right. PC cut off point 90%. Scale: variance divided by observation (nugget) variance.

same covariance model but happen to have extra noise in the observations (independent in every data) it can be of interest to see how this noise is propagated in the various predictions. In Figure 8 we show how a unit variance will show up as variance in the grid predictions. In reality such noise will have much less variance than the field itself, so the scale is misleading if the various plots are compared, but the non uniform result is clear. Here the principal components will remove the most pronounced peaks of the nugget variance in grid points.

Acknowledgement: The proof that $1/(1+r)$ is positive definite in was given some years ago by Associated Professor Arne Enqvist, Linkoping University, the day after I had questioned if that could be proved. Perhaps it can be found somewhere else in the litterature but I have not come across it.

6 References

Cressie, N. and Huang, H-C. (1999+). Classes of nonseparable, spatio-temporal stationary covariance functios. To appear in JASA.

Jones, R. H. and Zhang, Y. (1997). Models for continuous stationary space-time processes. In Modelling Longitudinal and Spatially Correlated Data, Editors Gregorie et al, Springer Lecture Notes in Statistics, Nr 122, 289–298.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1997). Multivariate Analysis, Academic Press, 6th printing.

Matérn, B. (1986). Spatial Variation. Springer Lecture Notes in Statistics, No 6. (First edition published 1960).

Nordgaard, A. and Hjorth, U. (1993). Statistical extrapolation of nutrient concentrations in the Baltic Sea. Environmetrics 4, 279–309.

Whittle, P., (1962). Topographic correlation, power-law covariance functions, and diffusion. Biometrika, 40, 305–314.