ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/24065345

The Kriged Kalman Filter

Article *in* Test · February 1998

DOI: 10.1007/BF02565111 · Source: RePEc

CITATIONS
189

READS
489

4 authors, including:

Colin R Goodall
AT&T
32 PUBLICATIONS **1,908** CITATIONS

SEE PROFILE

F. J. Alonso
University of Granada
22 PUBLICATIONS **388** CITATIONS

SEE PROFILE

All content following this page was uploaded by Colin R Goodall on 02 April 2015.

The user has requested enhancement of the downloaded file.

The Kriged Kalman filter

Kanti V. Mardia
Department of Statistics, University of Leeds
Leeds LS2 9JT, U.K. (K.V.Mardia@leeds.ac.uk)
Colin Goodall
Risk Adjustment Inc, Box 640
Rumson, NJ 07760, U.S.A.,(cgoodall@monmouth.com)
Edwin J. Redfern
Department of Statistics, University of Leeds
Leeds LS2 9JT, U.K. (E.J.Redfern@leeds.ac.uk)
Francisco J. Alonso
Department of Statistics and O.R., University of Granada
Granada 18071, Spain (falonso@goliat.ugr.es)

Abstract

In recent years there has been growing interest in spatial-temporal modelling, partly due to the potential of large scale data in pollution and global climate monitoring to answer important environmental questions. We consider the Kriged Kalman filter (KKF), a powerful modelling strategy which combines the two well-established approaches of (a) Kriging, in the field of spatial statistics, and (b) the Kalman filter, in general state space formulations of multivariate time series analysis. We give a brief introduction to the model and describe its various properties, and highlight that the model allows prediction in time as well as in space, simultaneously. Some special cases of the time series model are considered. We give some procedures to implement the model, also illustrated through a practical example. The paper concludes with a discussion.

Key Words: Bending energy, EM algorithm, Kalman filter, Karahunen-Loeve expansions, Kriging, pollution, spatial temporal modelling, state-space model.
AMS subject classification: 62M10, 62M30, 62M99.

1 Introduction

Spatial-temporal modelling has developed through applications in Geostatistics, Hydrology (Rouhani and Myers, 1990), Meteorology (Haslett, 1989) and more recently Environmental pollution monitoring (Guttorp and Sampson, 1994; Mardia and Goodall, 1993).

Goodall and Mardia (1994) outline an approach based on the spatial temporal general state space model designed to model the evolution of

spatial fields through time. Estimation is recursive, using the Kalman filter. This approach will form the basis of our space time modelling of the data in this paper.

The approach combines Kriging and the Kalman filter. We call it the Kriged Kalman filter (KKF). Recall that

(a) Kriging has provided a successful Weiner prediction approach in Spatial Statistics/Geostatistics.

(b) The Kalman filter gives a well-established recursive procedure for estimation in general state space models applied to time series.

We summarize each of Kriging and Kalman filter separately as an introduction to this paper.

Kriging. Let $\{x(\mathbf{s})\}$ be a random field $\mathbf{s} \in \Re^k$. The simplest case is that $y(\mathbf{s}) = x(\mathbf{s}) - f(\mathbf{s})$ is second order stationary where $f(\mathbf{s})$ is a trend function. Given the covariance function $\sigma(\mathbf{s}, \mathbf{s}')$ of $y(\mathbf{s})$, we can predict $x(\mathbf{s})$ at a new site $\mathbf{s} = \mathbf{s}_0$. The unbiased linear prediction of $x(\mathbf{s})$ is called Kriging. Note that prediction extends to intrinsic random fields where the field is incrementally stationary in space. Cressie (1993), 2nd edition, and Ripley (1981) provide a good introduction to the subject.

Kalman filter. Consider the state space model with the following scalar *observation* equation and p-dimensional *system* equations

$$x(t) = \mathbf{h}^T \boldsymbol{\alpha}(t) + \varepsilon(t),$$
$$\boldsymbol{\alpha}(t) = P\boldsymbol{\alpha}(t-1) + K\boldsymbol{\eta}(t)\,,$$

where $\mathbf{h}$ is the parameter p-vector, $\boldsymbol{\alpha}(t)$ is the state p-vector, $\varepsilon(t)$ is the scalar observation error, $P : p \times p$ is the transition matrix, $K : p \times d$ is the innovation coefficient matrix, and $\boldsymbol{\eta}(t)$ is the innovation (or system error or state noise) d-vector. The state space model is a latent model where $\boldsymbol{\alpha}(t)$ is of interest but $x(t)$ is observed. Predictions of $x(t)$ are also important. The Kalman filter provides recursive predictions, that is, given $x(1), \ldots, x(t-1)$, we can estimate $\widehat{x}(t)$, as well as the state vectors $\boldsymbol{\alpha}(1), ..., \boldsymbol{\alpha}(t-1)$. On observing $x(t)$ we can update our estimate of $\boldsymbol{\alpha}(t-1)$, by the product of the prediction error $x(t) - \widehat{x}(t)$ and the Kalman gain, to estimate $\boldsymbol{\alpha}(t)$. Maximum-likelihood estimation can be used for any unknown parameters. For a straightforward introduction, see Meinhold and Singpurwalla (1983),

and for a comprehensive study, see West and Harrison (1989). The state space model in (1.1) with scalar $x(t)$ generalizes readily to the multivarate state space model with n-vector $\mathbf{x}(t)$, and parameter *matrix* $H : n \times p$.

A review of spatio-temporal modelling up to 1993 has been provided by Goodall and Mardia (1994). Recent new model-based work includes Diggle (1998), Smith (1998), Wikle and Cressie (1997), Wikle *et al* (1998), and various papers in the edited volume by Gregoire *et al* (1997). Wikle and Cressie (1997) provide one important implementation of this model through empirical orthogonal functions (EOFs) which was suggested in Goodall and Mardia (1994). Smith's (1998) work extends some aspects of Mardia and Goodall (1993). Diggle's (1998) model places the spatial component in the system equation (cf. Green and Titterington, 1988). Here we provide various further properties of the KKF. Also, different implementations are aired. In particular, we give a specific implementation using pollution data.

2 The KKF model

The KKF model is a particular type of General State Space (GSS) model for Spatial-Temporal data (ST-GSS model), as defined by Goodall and Mardia (1994). In its simplest form the ST-GSS model is as follows. The spatial-temporal field

$$x(\mathbf{s}, t),\ \mathbf{s} \in S \subset \Re^k,\ t \in T \subset \Re \tag{2.1}$$

is decomposed into mean and error components

$$x(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \epsilon(\mathbf{s}, t)\,. \tag{2.2}$$

The mean component is expressed as a time-varying linear combination $\boldsymbol{\alpha}(t)$ of spatial fields $\mathbf{h}(\mathbf{s})$, which we call the *common fields* of the ST-GSS model.

$$\mu(\mathbf{s}, t) = h_1(\mathbf{s})\alpha_1(t) + h_2(\mathbf{s})\alpha_2(t) + \ldots + h_p(\mathbf{s})\alpha_p(t) = \mathbf{h}(\mathbf{s})^T\boldsymbol{\alpha}(t)\,. \tag{2.3}$$

(We remark on the case of slowly varying spatial fields, $\mathbf{h}(\mathbf{s}, t)$, in Goodall and Mardia (1994).) The error component $\varepsilon(\mathbf{s}, t)$ is a zero-mean spatial-temporal field. At each spatial location $\mathbf{s}$, the substitution of (2.3) into (2.2) gives the observation equation of the ST-GSS model.

A decomposition such as (2.3) always exists, e.g. the Karahunen-Loeve expansion. Cohen and Jones (1969), and Creutin and Obled (1982), discuss decomposition of $x(\mathbf{s}, t)$ into a weighted sum of EOFs. They discuss principal components decomposition of the $T \times n$ matrix X, comprising a time series of length T with observations on n sites at each occasion, and subsequent interpolation of the principal n-vectors to obtain spatial fields, the EOFs. For $p < \min(T, n)$ the spatial-temporal field $\varepsilon(\mathbf{s}, t)$ represents, in the empirical setting (of $X : T \times n$), the residuals from the rank-p SVD.

The vector $\boldsymbol{\alpha}(t)$ represents the *state* of the system of (2.3) for $\mu(\mathbf{s}, t)$. The state equations

$$\boldsymbol{\alpha}(t) = P\boldsymbol{\alpha}(t-1) + K\boldsymbol{\eta}(t) \tag{2.4}$$

introduce a stochastic component into the trend $\mu(\mathbf{s}, t)$. The innovation at time t is $K\boldsymbol{\eta}(t)$, where we assume $\boldsymbol{\eta}(t) \sim NID(0, \Sigma_{\boldsymbol{\eta}})$. Generally, some but not all of the parameters P, K, and $\Sigma_{\boldsymbol{\eta}}$ are specified by the model, so known.

Our choice of $h_j(\mathbf{s})$, $j = 1, ..., p$, is driven by a priori modelling considerations and also by the data. The p common fields $h_j(\mathbf{s})$ divide into two sets: q *trend fields* and r *principal fields* ($p = q + r$). The trend fields typically are selected from the conventional choices in Kriging spatial data, including constant, linear and quadratic functions of coordinate dimensions. The principal fields are selected from a basis of the space of all possible spatial Kriging estimates for a given set of m "normative" sites, and for given second order spatial structure (covariance or variogram). There are up to m principal fields for the m normative sites. Normative sites and principal fields are discussed and defined in Section 3. The usual use of data for model estimation, selection and validation includes choosing the set of normative sites and second order structure.

The ST-GSS model, given by equations (2.2)–(2.4), has very general interest and applicability. With the objective of an unbiased prediction with minimum variances, fitting the ST-GSS model addresses: "To what extent can data of the spatial-temporal field $x(\mathbf{s}, t)$, be represented as a linear combination of common spatial fields ((2.3)), where the linear combination given by the state vector is updated from one occasion to the next by the linear transition matrix P with stochastic innovations?" A stronger result is to find that *all* temporal dependence in $X(\mathbf{s}, t)$ is conveyed by the state vector $\boldsymbol{\alpha}(t)$ and (2.4). The error component $\varepsilon(\mathbf{s}, t)$ is then a spatial-

correlated error process, with

$$\mathrm{cov}\,(\varepsilon(\mathbf{s},t),\varepsilon(\mathbf{s}',t')) = 0 \quad \text{for } t \neq t' \quad \text{all } \mathbf{s},\mathbf{s}'. \tag{2.5}$$

We say that (2.2)–(2.4) define the weak ST-GSS model, and equations (2.2)–(2.5) define the strong ST-GSS model, which we call the Kriged Kalman filter (KKF) model in this paper.

In the following sections we describe how to compute the principal fields to be used in the observation equation (Section 3), and various forms for the transition matrix P and other parameters in the observation and system equations (Section 4). We next turn to strategies for estimation (Section 5), to some general properties of KKF (Section 6), and to an extended example using air pollution data taken from in and around Leeds, United Kingdom (Section 7). The paper concludes with a discussion of some extensions to the ST-GSS and KKF models (Section 8).

3 Principal fields

The systematic spatial component in $x(\mathbf{s},t)$ is modeled as time-varying linear combinations of p common fields, comprising q trend fields and $r = p - q$ principal fields, written in order as

$$\mathbf{h}(\mathbf{s})^T = (h_1(\mathbf{s}),\ldots,h_q(\mathbf{s}),h_{q+1}(\mathbf{s}),\ldots,h_{q+r}(\mathbf{s}))\,.$$

The fields are found by considering common trend components, and common spatial dependency across time. Consider the spatial linear model

$$x(\mathbf{s}) = \mathbf{f}(\mathbf{s})^T\boldsymbol{\beta} + \zeta(\mathbf{s}) \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_1,...,\beta_q)^T$, $\mathbf{f}(\mathbf{s}) = (f_1(\mathbf{s}),...,f_q(\mathbf{s}))^T$, and the $f_j(\mathbf{s})$ ($j = 1,\ldots,q$) are trend fields of given polynomial form in the co-ordinates of $\mathbf{s}$. Let

$$\mathrm{cov}\,(\zeta(\mathbf{s}),\zeta(\mathbf{s}')) = \sigma_\zeta(\mathbf{s},\mathbf{s}')\,, \tag{3.2}$$

a conditionally positive-definite function. Suppose, hypothetically, that observations $\mathbf{x} = (x_1,\ldots,x_m)$ are taken at the m normative sites, $m \geq p$, $\mathbf{s}_1^*,\ldots,\mathbf{s}_m^*$. The common spatial dependency across time is expressed as the set of Kriging predictors, possibly constrained to be less than full rank (see below), for sets of observations at these normative sites. The normative

sites will typically coincide, but need not coincide, with sites at which actual observations are made.

Write $\boldsymbol{\sigma}_\zeta(\mathbf{s})$ for the m-vector of covariances with i'th element $\sigma_\zeta(\mathbf{s}_i^*, \mathbf{s})$, and write Σ_ζ for the $m \times m$ covariance matrix with $(\Sigma_\zeta)_{ij} = \sigma_\zeta(\mathbf{s}_i^*, \mathbf{s}_j^*)$. Let $\mathbf{f}_j$ be the m-vector of the j'th trend field at the normative sites, with i'th element $f_j(\mathbf{s}_i^*)$.

The Kriging predictor is

$$\widehat{x}(\mathbf{s}) = \mathbf{f}(\mathbf{s})^T A\mathbf{x} + \boldsymbol{\sigma}_\zeta(\mathbf{s})^T B\mathbf{x} \tag{3.3}$$

where, writing $F : m \times q$ for the trend matrix with ij'th element $f_j(\mathbf{s}_i)$,

$$A = (F^T {\Sigma_\zeta}^{-1} F)^{-1} F^T {\Sigma_\zeta}^{-1} \tag{3.4}$$

is the matrix of trend, and

$$B = {\Sigma_\zeta}^{-1} - {\Sigma_\zeta}^{-1} F (F^T {\Sigma_\zeta}^{-1} F)^{-1} F^T {\Sigma_\zeta}^{-1} \tag{3.5}$$

is the partial information matrix, or generalized bending energy matrix where we assume Σ_ζ and $F^T {\Sigma_\zeta}^{-1} F$ to be non-singular.

Assume, for the purposes of discussion, that the columns of F are linearly independent. Consider the spectral decomposition of B,

$$B = UDU^T\,, \qquad B\mathbf{u}_i = d_i \mathbf{u}_i\,, \tag{3.6}$$

where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ and $D = \operatorname{diag}\{d_1, \ldots, d_m\}$. At least q of the eigenvalues d_i are equal to zero; to simplify the following exposition we assume that exactly q eigenvalues equal zero. Write the eigenvalues in non-decreasing order, $0 = d_1 = d_2 = \ldots = d_q < d_{q+1} \leq d_{q+2} \leq \ldots \leq d_m$. The generalized bending energy matrix B has maximum rank, $m - q$, and the vectors of B are orthogonal to the columns of F, $BF = 0$, *i.e.* $\mathbf{f}_j^T \mathbf{u}_i = 0$ for $j = 1, \ldots, q$ and $i = q+1, \ldots, m$.

Suppose that observations taken at the normative sites coincide with the i'th eigenvector of B, $i = 1, \ldots, m$. The Kriging predictor (3.3) is

$$\begin{aligned} \widehat{x}(\mathbf{s})_{\mathbf{x}=\mathbf{u}_i} &= \mathbf{f}(\mathbf{s})^T A\mathbf{u}_i + \boldsymbol{\sigma}_\zeta(\mathbf{s})^T B\mathbf{u}_i \\ &= \mathbf{f}(\mathbf{s})^T A\mathbf{u}_i + d_i \boldsymbol{\sigma}_\zeta(\mathbf{s})^T \mathbf{u}_i\,. \end{aligned} \tag{3.7}$$

Any m-vector of observations $\mathbf{x}$ is the linear combination $U^T\mathbf{x}$ of the $\mathbf{u}_i$, and the Kriging predictor $\widehat{x}(\mathbf{s})$ is the linear combination $U^T\mathbf{x}$ of the $\widehat{x}(\mathbf{s})_{\mathbf{x}=\mathbf{u}_i}$

in (3.7). Thus $\widehat{x}(\mathbf{s})$ is a linear combination of the q trend fields $f_j(\mathbf{s})$ $(j = 1, \ldots, q)$ and the $m - q$ *principal fields* $\boldsymbol{\sigma}_\zeta(\mathbf{s})^T \mathbf{u}_i$ $(i = q + 1, \ldots, m)$, evaluated at $\mathbf{s}$. These functions span the space of all Kriging solutions with observations at the m given sites, and the specified trend fields and covariance.

In practice the r principal fields used in the KKF model may include either all $(r = m - q)$ or a subset $(r < m - q)$ of the principal fields. Then

$$p = q + r \leq m\,.$$

Bookstein (1989) notes, for thin plates splines, the association of small eigenvalues of the bending energy matrix with large-scale variation ("global features") and of large eigenvalues with patterns of local variation. Let $\{j_1, \ldots, j_r\}$ denote a subset of size $r \leq m - q$ of $\{q + 1, \ldots, m\}$. Then the mean component in the observation equation of the KKF model has the form

$$\mu(\mathbf{s}, t) = \sum_{j=1}^{q} h_j(\mathbf{s})\alpha_j(t) + \sum_{k=1}^{r} h_{q+k}(\mathbf{s})\alpha_{q+k}(t)\,, \tag{3.8}$$

where

$$h_j(\mathbf{s}) = f_j(\mathbf{s}), \qquad j = 1, \ldots, q \tag{3.9}$$

$$h_{q+k}(\mathbf{s}) = \boldsymbol{\sigma}_\zeta(\mathbf{s})^T \mathbf{u}_{j_k}, \qquad k = 1, \ldots, r\,. \tag{3.10}$$

When the full set of principal fields with non-zero eigenvalues is used, $r = m - q$, then we write $j_k = q + k$ in (3.10).

The Kriging predictor depends on the covariogram, the spatial pattern of the points, and the trend fields in the F matrix. Thus for the KKF we need to develop methods for determining a suitable, possibly optimal, set of normative sites, while choosing and fitting an appropriate covariogram model, identifying an F matrix, and choosing all or a subset of the principal fields.

4 Specifying the temporal component

In this section we consider various specifications of the temporal component of the ST-GSS model. This takes us some way to specifying the $p \times p$ transition matrix P of the system equations, so that, for particular choices

of temporal component, P may be fixed or P may have a few parameters in place of the nominal set of p^2 elements. We discuss two approaches, namely autoregressive modelling, and dynamic linear modelling through an augmented Holt-Winters model. The multivariate setting adds considerably to the complexity of model specification and selection, as in principle it can associate a separate univariate time-series model with each common field, and include additional cross-correlation terms. In our applications we typically repeat a single structure for each common field and take the cross-correlation terms to be zero.

4.1 Autoregressive specification

The standard state-space form of the autoregressive model includes structural zeros in the parameter vector of the observation equation. These correspond to elements of the state vector, that while essential for updating, do not directly affect the mean component. Write, for each $\mathbf{s}$,

$$x(\mathbf{s},t) = \mathbf{h}^{\dagger}(\mathbf{s})^T \boldsymbol{\alpha}(t) + \varepsilon(\mathbf{s},t)\,, \tag{4.1}$$

where for an autoregressive model of order u, AR(u),

$$\mathbf{h}^{\dagger}(\mathbf{s})^T = (h_1(\mathbf{s}), \mathbf{0}_{u-1}^T, h_2(\mathbf{s}), \mathbf{0}_{u-1}^T, \ldots, h_{q+r}(\mathbf{s}), \mathbf{0}_{u-1}^T)\,, \tag{4.2}$$

an $(q+r)u$ vector, denoting by $\mathbf{0}_{u-1}^T$ a row vector of $(u-1)$ zeros. Note that the simplest AR(u) model would have a single vector of parameters, $\mathbf{h}^{\dagger}(\mathbf{s})^T = (h_1(\mathbf{s}), \mathbf{0}_{u-1}^T)$, length u.

The non-zero fields $h_1(\mathbf{s}), \ldots, h_{q+r}(\mathbf{s})$ comprise q trend fields and r principal fields. The quantity p denotes the dimension of the state vector of the GSS model, that is, the lengths of $\mathbf{h}^{\dagger}(\mathbf{s})$ and $\boldsymbol{\alpha}(t)$, and $p = (q+r)u$. With reference to (2.3), we say that there are p common fields, comprising q trend fields, r principal fields, and $p_0 = (q+r)(u-1)$ zero fields, with $p = q + r + p_0$.

The transition matrix P of the state equation is $p \times p$ block diagonal, with $(q+r)$ identical blocks, each $u \times u$ and written Q_u:

$$P = \text{blockdiag}\,\{Q_u, \ldots, Q_u\} \tag{4.3}$$

where

$$Q_u = \begin{bmatrix} \phi_p & 1 & 0 & \dots & 0 \\ \phi_{p-1} & 0 & 1 & \dots & 0 \\ \phi_{p-2} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_2 & 0 & 0 & \dots & 1 \\ \phi_1 & 0 & 0 & \dots & 0 \end{bmatrix}. \tag{4.4}$$

When $u = 1$, the AR(1) model, then $p_0 = 0$, $Q_u = [\phi]$ and

$$P = \text{diag}\{\phi, \dots, \phi\} = \phi I_{q+r}. \tag{4.5}$$

The transition matrix has a repetitive block diagonal form; this is not true when the components associated with each common field follow different autoregressive models. Then we may have, for example, when $u = 1$,

$$P = \text{diag}\{\phi_1, \dots, \phi_{q+r}\}. \tag{4.6}$$

Note that this state-space form of the autoregressive model is not unique but is minimal (Akaike, 1973).

4.2 Dynamic linear model specification

An alternative model is one based on the dynamic linear model (DLM) (Harrison and Stevens, 1976). The simplest DLM is the Holt-Winters model, based on slope and level. Consider the ST-GSS model with $\mathbf{h}^\dagger(\mathbf{s})$ the same as for AR($u = 2$),

$$\mathbf{h}^\dagger(\mathbf{s})^T = (h_1(\mathbf{s}), 0, h_2(\mathbf{s}), 0, \dots, h_{q+r}(\mathbf{s}), 0). \tag{4.7}$$

The transition matrix P is $p \times p$, $p = 2(q + r)$,

$$P = \text{blockdiag}\{Q_2^*, \dots, Q_2^*\} \tag{4.8}$$

where

$$Q_2^* = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

This completely specifies the P matrix, reducing the number of parameters to be estimated.

For this Holt-Winters model,

$$\mu(\mathbf{s},t) = h_1(\mathbf{s})\alpha_1(t)+h_2(\mathbf{s})\alpha_3(t)+\ldots+h_j(\mathbf{s})\alpha_{2j-1}(t)+\ldots+h_{q+r}(\mathbf{s})\alpha_{p-1}(t) \quad (4.9)$$

where each of $\alpha_1(t), \alpha_3(t), \ldots, \alpha_{p-1}(t)$ follows a linear trajectory in the mean. Write $\mathbf{k}_{2j-1}$ and $\mathbf{k}_{2j}$ for the $(2j-1)$'th and $(2j)$'th rows of the innovation parameter matrix K. Then

$$\begin{aligned} \alpha_{2j-1}(t) &= \alpha_{2j-1}(t-1) + \alpha_{2j}(t-1) + \mathbf{k}_{2j-1}\boldsymbol{\eta}(t) \\ \alpha_{2j}(t) &= \alpha_{2j}(t-1) + \mathbf{k}_{2j}\boldsymbol{\eta}(t) \end{aligned} \quad (4.10)$$

That is, $\alpha_{2j-1}(t)$, the scalar multiplier of $h_j(\mathbf{s})$, increases by $\alpha_{2j}(t-1)$ each unit of time up to stochastic innovations, with level $\alpha_{2j-1}(1)$ at time $t = 1$.

An interesting alternative is to incorporate separate spatial fields for level $h_L(\mathbf{s})$ and slope $h_S(\mathbf{s})$. With $p = 3$ common fields, including one zero field, $\mathbf{h}^\dagger(\mathbf{s})^T = (h_L(\mathbf{s}), h_S(\mathbf{s}), 0)$, and $\boldsymbol{\alpha}(t)^T = (\alpha_L(t), \alpha_S(t), \alpha'_S(t))$, the mean is

$$\mu(\mathbf{s},t) = h_L(\mathbf{s})\alpha_L(t) + h_S(\mathbf{s})\alpha_S(t), \quad (4.11)$$

and the 3×3 transition matrix is

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \quad (4.12)$$

The "base" spatial field is the random multiple $\alpha_L(t)$ of $h_L(\mathbf{s})$, where $\alpha_L(t)$ follows Brownian motion centered at $\alpha_L(1)$. The contribution from the "slope" is the random multiple $\alpha_S(t)$ of $h_S(\mathbf{s})$, where $\alpha_S(t)$ follows its own scalar "level plus slope" model with base value $\alpha_S(1)$ and slope $\alpha'_S(t)$.

This pattern may be repeated to include several sets of 3 state parameters, and fields possibly related to trend and principal fields. However, for the case above with $p = 3$, neither of the two common fields $h_L(\mathbf{s})$ and $h_S(\mathbf{s})$ will be either a trend field or a principal field, except in very special cases. Instead, the model may express $h_L(\mathbf{s})$ and $h_S(\mathbf{s})$ as linear combinations of the trend and principal fields. The coefficients of the linear combinations become parameters in the observation equation. Write these fields as

$$\begin{aligned} \mathbf{h}(\mathbf{s})^T &= (h_1(\mathbf{s}), h_2(\mathbf{s}), \ldots, h_{q+r}(\mathbf{s})) \\ h_L(\mathbf{s}) &= \boldsymbol{\gamma}_L^T\mathbf{h}(\mathbf{s}) \quad (4.13) \\ h_S(\mathbf{s}) &= \boldsymbol{\gamma}_S^T\mathbf{h}(\mathbf{s}). \quad (4.14) \end{aligned}$$

The mean field at time t can be written $\mu(\mathbf{s}, t) = \mathbf{h}(\mathbf{s})^T \boldsymbol{\alpha}^+(t)$ where

$$\boldsymbol{\alpha}^+(t) = \alpha_L(t)\boldsymbol{\gamma}_L + \alpha_S(t)\boldsymbol{\gamma}_S.$$

Note that the above is essentially an ARIMA(0,2,2) model, a Holt-Winters model with the additional term comprising the base level $h_L(\mathbf{s})$.

5 Estimation

Our discussion of estimation for the KKF is in several parts. In the first part, Section 5.1, we review the Kalman filter and its use in estimation of the states $\boldsymbol{\alpha}(t)$ given the data, assuming the parameters and the common fields to be specified. In the second part, Section 5.2, we describe maximum-likelihood estimation (MLE), and its implementation using Newton-Raphson and the EM algorithm, of the parameters P, $\Sigma_{\boldsymbol{\eta}}$, and $\Sigma_{\boldsymbol{\varepsilon}}$, given the common fields, matrix H. In the third part, Section 5.3, we discuss implementation of the full algorithm, which incorporates both spatial modelling, to define the set of common fields leading to H, and MLE and the Kalman filter. We propose a strategy for iterating between the spatial modelling and multivariate time series modelling steps of the full algorithm.

5.1 The Kalman filter recursion

Let us now assume that we have observations at the n sites, $\mathbf{s}_i$, $i = 1, \ldots, n$, on T occasions, $t = 1, \ldots, T$. Write the t'th observation vector as

$$\mathbf{x}_t^T = (x(\mathbf{s}_1, t), \ldots, x(\mathbf{s}_n, t)) \ , \tag{5.1}$$

and collect these into the $T \times n$ matrix of observations denoted X with t'th row $\mathbf{x}_t^T$.

Suppose that the $n \times p$ parameter matrix is

$$H = \begin{pmatrix} \mathbf{h}^T(\mathbf{s}_1) \\ \vdots \\ \mathbf{h}^T(\mathbf{s}_n) \end{pmatrix}. \tag{5.2}$$

The observation equation is

$$\mathbf{x}_t = H\boldsymbol{\alpha}(t) + \boldsymbol{\varepsilon}(t) \tag{5.3}$$

with the usual system equation (2.4).

The KKF model includes spatial parameters in the covariogram model and temporal parameters in the GSS model. Were the spatial parameters known, then maximum-likelihood estimators of the temporal parameters can be obtained using established Kalman filter techniques.

Let $\mathbf{a}_{t-1|t-1}$, $\mathbf{a}_{t|t-1}$, and $\mathbf{a}_{t|t}$ be estimates of $\boldsymbol{\alpha}(t-1)$, $\boldsymbol{\alpha}(t)$, and $\boldsymbol{\alpha}(t)$ based on observations available to times $t-1$, $t-1$, and t respectively. Let $C_{t-1|t-1}$, $C_{t|t-1}$, and $C_{t|t}$ be the covariances of $\mathbf{a}_{t-1|t-1}$, $\mathbf{a}_{t|t-1}$, and $\mathbf{a}_{t|t}$, respectively.

Given $\mathbf{a}_{t-1|t-1}$ and $C_{t-1|t-1}$, and using the system equation, the prediction equations for the state of the Kalman filter are

$$\mathbf{a}_{t|t-1} = P\mathbf{a}_{t-1|t-1} \tag{5.4}$$

$$C_{t|t-1} = PC_{t-1|t-1}P^T + K\Sigma_{\boldsymbol{\eta}}K^T\,. \tag{5.5}$$

Write $\widehat{\mathbf{x}}_{t|t-1}$ for the plug-in estimate of the observation at t. The one-step prediction error at time t is

$$\mathbf{e}_t = \mathbf{x}_t - \widehat{\mathbf{x}}_{t|t-1}\,, \tag{5.6}$$

and write $\mathrm{cov}\,(\mathbf{e}_t) = D_{t|t-1}$ for the one step prediction error variance. ¿From the observation equation,

$$\widehat{\mathbf{x}}_{t|t-1} = H\mathbf{a}_{t|t-1} \tag{5.7}$$

$$D_{t|t-1} = HC_{t|t-1}H^T + \Sigma_{\boldsymbol{\varepsilon}}\,. \tag{5.8}$$

The Kalman gain matrix at time t is defined to be

$$G_t = C_{t|t-1}H^T {D_{t|t-1}}^{-1} \tag{5.9}$$

$$= C_{t|t-1}H^T(HC_{t|t-1}H^T + \Sigma_{\boldsymbol{\varepsilon}})^{-1}\,. \tag{5.10}$$

Then the smoothed estimates of the state $\mathbf{a}_{t|t}$ and its covariance matrix $C_{t|t}$ are given by the Kalman updating equations

$$\mathbf{a}_{t|t} = \mathbf{a}_{t|t-1} + G_t\mathbf{e}_t = \mathbf{a}_{t|t-1} + G_t(\mathbf{x}_t - H\mathbf{a}_{t|t-1})\,, \tag{5.11}$$

$$C_{t|t} = C_{t|t-1} - G_tD_{t|t-1}G_t = C_{t|t-1} - G_tHC_{t|t-1}\,. \tag{5.12}$$

The set of equations, (5.4)–(5.12) requires initial estimates $\mathbf{a}_{0|0}$ and $C_{0|0}$.

Thus given data $\mathbf{x}_1, \ldots, \mathbf{x}_T$, an initial estimate $\mathbf{a}_{0|0}$, the true covariance $C_{0|0}$, the parameter matrix H, and parameters P, $\Sigma_{\boldsymbol{\varepsilon}}$, and $\Sigma_{\boldsymbol{\eta}}$, the Kalman filter recursion provides estimates $\mathbf{a}_{1|1}, \ldots, \mathbf{a}_{T|T}$ of $\boldsymbol{\alpha}(1), \ldots, \boldsymbol{\alpha}(T)$, and the covariances of the estimates, $C_{1|1}, \ldots, C_{T|T}$. Often we take $C_{0|0} = 0$.

In practice, the entire data matrix X is available to us at once, and the use of the Kalman filter (forward) recursion is seen to be a computational device. Estimates of the $\boldsymbol{\alpha}(t)$ that make use of all the data, instead of only the data up to time t, are obtained when the Kalman recursion is followed by Kalman back-recursion, possibly iterating the forward and back recursions one or more times. Detailed formulae for the back recursion can be found in standard textbooks.

5.2 Maximum likelihood estimation

The likelihood uses the prediction error $\mathbf{e}_t$ from (5.6). The prediction error at spatial location $\mathbf{s}$ at time t can be written

$$e(\mathbf{s}, t) = \mathbf{h}(\mathbf{s})^T(\boldsymbol{\alpha}(t) - \mathbf{a}_{t|t-1}) + \varepsilon(\mathbf{s}, t)\,.$$

Further, we condition the likelihood on $\mathbf{a}_{0|0}$ and $C_{0|0} = 0$. Conditional on $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}$, the $e(\mathbf{s}, t)$ have zero mean and covariance matrix $D_{t|t-1}$. Write $\boldsymbol{\theta}$ for the unknown parameters in P, $\Sigma_{\boldsymbol{\varepsilon}}$, and $\Sigma_{\boldsymbol{\eta}}$. The likelihood for $\boldsymbol{\theta}$ is

$$\ln L(X; \boldsymbol{\theta}) = -\tfrac{1}{2} \sum_{t=1}^{T} \ln \mid D_{t|t-1} \mid - \tfrac{1}{2} \sum_{t=1}^{T} \mathbf{e}_t^T {D_{t|t-1}}^{-1} \mathbf{e}_t, \tag{5.13}$$

which is a highly non-linear function of the unknown parameters.

A basic numerical technique is to fix the parameters $\boldsymbol{\theta}$ at some initial values and use the Kalman filter recursions to compute the $\widehat{\mathbf{x}}_{t|t-1}$ and the $D_{t|t-1}$ and thus the log-likelihood, given by (5.13). A series of Kalman filter forward and back recursions, and computation of $\ln L$, comprises the inner step in a numerical optimization algorithm. Newton-Raphson and EM algorithms have been used in the literature.

In the Newton–Raphson approach, one fixes the initial $\mathbf{a}_{0|0}$, and then develops a set of recursions for the log-likelihood and its first two derivatives. See Gupta and Mehra (1974), Ansley and Kohn (1984), Jones (1980), and Janacek and Swift (1993).

We now describe the approach using the EM algorithm following Shumway and Stoffer (1982) and Watson and Engle (1983). Suppose that the initial estimate $\mathbf{a}_{0|0}$ is drawn from a Gaussian$(\boldsymbol{\mu}_0, \Sigma_0)$ distribution, and $\boldsymbol{\varepsilon}(t)$ and $\boldsymbol{\eta}(t)$ are jointly Gaussian. The complete data is taken to be $(\mathbf{a}_{0|0}, \boldsymbol{\alpha}(1), \ldots, \boldsymbol{\alpha}(T), \mathbf{x}_1, \ldots, \mathbf{x}_T)$, but of course the $\boldsymbol{\alpha}(t)$ are not observed. Suppose that the innovation parameter matrix K equals the identity, $K = I_p$. The likelihood of the complete data can be written

$$\begin{aligned}
\ln L = \quad & -\tfrac{1}{2}\ln \mid \Sigma_0 \mid - \quad \tfrac{1}{2}(\mathbf{a}_{0|0} - \boldsymbol{\mu}_0)^T {\Sigma_0}^{-1}(\mathbf{a}_{0|0} - \boldsymbol{\mu}_0) \\
& -\tfrac{T}{2}\ln \mid \Sigma_{\boldsymbol{\eta}} \mid - \quad \tfrac{1}{2}\sum_{t=1}^{T}(\boldsymbol{\alpha}(t) - P\boldsymbol{\alpha}(t-1))^T {\Sigma_{\boldsymbol{\eta}}}^{-1}(\boldsymbol{\alpha}(t) - P\boldsymbol{\alpha}(t-1)) \\
& -\tfrac{T}{2}\ln \mid \Sigma_{\boldsymbol{\varepsilon}} \mid - \quad \tfrac{1}{2}\sum_{t=1}^{T}(\mathbf{x}_t - H\boldsymbol{\alpha}(t))^T {\Sigma_{\boldsymbol{\varepsilon}}}^{-1}(\mathbf{x}_t - H\boldsymbol{\alpha}(t))\,. \qquad (5.14)
\end{aligned}$$

Our aim is to maximize $\ln L$ with respect to $\boldsymbol{\mu}_0$, Σ_0, P, $\Sigma_{\boldsymbol{\eta}}$, and $\Sigma_{\boldsymbol{\varepsilon}}$. We apply the EM algorithm conditionally with respect to the observed series $\mathbf{x}_1, \ldots, \mathbf{x}_T$ to maximize (5.14) where $\boldsymbol{\alpha}(t)$ is missing. Dempster *et al.* (1977) show that this results in a sequence of non-decreasing likelihoods. One difficulty is that there is a single composite term in (5.14) in $\boldsymbol{\mu}_0$ and Σ_0, so both cannot be estimated. Two alternatives are to (a) fix Σ_0 and estimate $\boldsymbol{\mu}_0$ and (2) to regard $\boldsymbol{\mu}_0$ as fixed and set $\Sigma_0 = 0$. This problem is discussed by Harvey (1989) and was recognized by Watson and Engle (1983). Estimation in the time domain then gives estimates of the parameters P, $\Sigma_{\boldsymbol{\eta}}$, $\Sigma_{\boldsymbol{\varepsilon}}$, and the states $\boldsymbol{\alpha}(t)$.

Accommodating missing values. Both $\mathbf{a}_{t|t}$ and $C_{t|t}$ are computed using the t'th observation $\mathbf{x}_t$. When $\mathbf{x}_t$ contains missing values, we omit those rows in H and compute $\mathbf{a}_{t|t}$ and $C_{t|t}$ relying on the non-missing data. An adjustment to the Kalman filter recursion formulae, and to the likelihood, is needed. This is in line with the usual procedure on missing values in Kalman filter applications (Jones 1980).

5.3 Implementation of the full algorithm

The full algorithm includes estimation of a set of common fields along with the Kalman filter recursions and parameter estimation discussed in Section 5.1 and Section 5.2. We first review some overall strategic concerns,

and then turn to specifics of the implementation of the full algorithm that incorporates both spatial and multivariate temporal modelling.

The objective in choosing the common fields is to capture as much as possible of the spatial variation at each occasion, and across occasions, as a linear combination of the common fields. These linear combinations, or state vectors, are constrained by the state equation of the ST-GSS, that is, of the KKF.

The common fields comprise trend fields and principal fields. Some alternative choices for the set of trend fields are constant only, linear in all spatial dimensions, linear in only some spatial dimensions, quadratic, etc. The principal fields come from a covariogram model, the Σ_ζ of Section 3, which is fit to the detrended data at all sites, but based on a selection of normative sites.

Given the choice of common fields, we choose the structure of a temporal model, possibly of a family of related temporal models, and proceed to estimate parameters for a selected model. This also leads us to estimation of $\Sigma_{\boldsymbol{\varepsilon}}$, the observation error, summarized most appropriately for spatial prediction as a spatial covariogram (or variogram).

The two spatial covariance models are separate, as the covariogram for $\Sigma_{\boldsymbol{\varepsilon}}$ reflects "smaller" scale variation, that is intended to be temporal uncorrelated. This is distinct from the "global", temporal, patterns incorporated in Σ_ζ (the initial covariogram of detrended data), used in determining principal fields. Indeed we can, and do, take $\Sigma_{\boldsymbol{\varepsilon}}$ to be the identity, implying that the spatial structure is modeled solely through the common fields.

The following table lists the components of KKF modelling, including both model specification and selection (denoted S) and parameter estimation (E) steps. The table underscores the importance of model specification/selection, which follow at least in part from an exploratory approach.

The Kalman filter provides maximum-likelihood estimates of $\boldsymbol{\alpha}(t)$ for given $\boldsymbol{\alpha}(0)$, $\Sigma_{\boldsymbol{\varepsilon}}$, $\Sigma_{\boldsymbol{\eta}}$, P, K, and H. Given H, the EM algorithm or Newton Raphson approach provides estimates of all the time parameters. Table 1 shows us how the problem of estimating the spatial components precedes and ties in with these steps. See Green (1995) for a jump-diffusion approach to simultaneous model specification and parameter estimation.

	spatial modelling
S	number and degree of trend fields
S	functional form of the covariogram Σ_ζ for principal fields
E	principal field covariogram parameters
S	number of normative sites
E	locations of normative sites
S	number of principal fields and whether they correspond to small, or large, or all eigenvalues of B
	temporal modelling
S	choice of general state space model
E	general state space model parameters
S	functional form of the covariogram $\Sigma_{\boldsymbol{\varepsilon}}$ for spatial errors
E	spatial errors covariogram parameters

Table 1: *Components of KKF modelling*

The full likelihood

$$\ln L(\boldsymbol{\theta}, \boldsymbol{\phi} \mid x(\mathbf{s}, t), \quad s \in \{\mathbf{s}_1, \ldots \mathbf{s}_n\}, t \in \{1, \ldots, T\})$$

is a function of temporal parameters $\boldsymbol{\theta}$, as in Section 5.2, and spatial parameters $\boldsymbol{\phi}$. The spatial parameters come from the components of KKF modelling corresponding to the first six rows of Table 1. Given the spatial model specification, the spatial parameters include the parameters of the covariogram and possibly also the locations of the normative sites. For data $x(\mathbf{s}, t)$, $\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, $t \in \{1, \ldots, T\}$, given values of the temporal parameters, $\boldsymbol{\theta}$, we can maximize $\ln L$ with respect to $\boldsymbol{\phi}$ and given values of the spatial parameters, $\boldsymbol{\phi}$, we can maximize $\ln L$ with respect to $\boldsymbol{\theta}$. This suggests an iterative two stage estimation process.

Algorithm I.

Step 1. Propose an $n \times p$ matrix H, by estimating a covariogram for Σ_ζ from the data, and combining columns containing values of q trend fields at the $\mathbf{s}_i$ together with columns containing the first $r = p - q$ principal fields $\mathbf{u}_{q+j}$, $j = 1, \ldots, r$, obtained from the generalized bending energy matrix B in Section 3, with normative sites $\mathbf{s}_i^* = \mathbf{s}_i$, $i = 1, \ldots, m = n$.

Step 2. Estimate, using the Kalman filter in the EM algorithm, the parameters P, $\Sigma_{\boldsymbol{\varepsilon}}$, $\Sigma_{\boldsymbol{\eta}}$ and the initial condition $\boldsymbol{\alpha}(0) \sim N(\boldsymbol{\mu}_0, \Sigma_0)$ (see above).

Step 3. Using the estimated parameters from the EM algorithm maximize the likelihood with respect to the covariogram parameters.

Repeat Step 2 and Step 3 until convergence is achieved.

The advantage of working in this manner, using a covariogram model, is that it allows spatial prediction at other spatial locations as well as extrapolation in time.

An alternative approach is to use the empirical covariance function in Step 1 and replace Step 3 by the following.

Algorithm II.

Step 3. Compute the covariance matrix Σ_ζ by substituting the estimated parameters into the formula for the covariance at time t for t large.

$$\begin{aligned} \operatorname{cov}\left(X(t,\mathbf{s}), X(t,\mathbf{s}')\right) &= \mathbf{h}(\mathbf{s})P^t\Sigma_0 P^{tT}\mathbf{h}(\mathbf{s}')^T \\ &+ \sum_{i=0}^{t-1}\mathbf{h}(\mathbf{s})P^i\Sigma_{\boldsymbol{\eta}}P^{iT}\mathbf{h}(\mathbf{s}')^T + \operatorname{cov}\left(\boldsymbol{\varepsilon}(\mathbf{s}), \boldsymbol{\varepsilon}(\mathbf{s}')\right). \end{aligned} \tag{5.15}$$

See (6.5) and the surrounding discussion in Section 6.2 for details. If the eigenvalues of P are less than unity, then the first term on the right of (5.15) is negligible when t is large. Note that the two covariances in the formula are respectively Σ_ζ and $\Sigma_{\boldsymbol{\varepsilon}}$.

Step 4. Compute the H matrix using the new covariance structure.

Repeat Steps 2-4 until convergence is achieved.

Algorithm II can be used for temporal prediction at the observed spatial locations, and can be used for spatial interpolation if a covariogram model for $\Sigma_{\boldsymbol{\varepsilon}}$ is assumed.

6 General properties

6.1 Spatial prediction for data at time t

Given data $\{\mathbf{x}_k; k = 1, \ldots, T\}$, prediction at time t, $t \leq T$, at unmonitored sites will be called *spatial prediction*. The straightforward approach here uses plug-in estimates in the conditional Gaussian distribution of the observations at the unmonitored sites given both the observed data and current estimate of the state. For a different approach, see Cressie and Wikle (1997).

Let $\mathbf{x}_t^*$ denote the random vector, and H^* the matrix of common field values, at n^* unmonitored sites. Let $\boldsymbol{\varepsilon}^*(t)$ denote the corresponding measurement error, and write $\mathrm{cov}\,(\boldsymbol{\varepsilon}(t), \boldsymbol{\varepsilon}^*(t)) = \Sigma_{\boldsymbol{\varepsilon}}^*$ and $\mathrm{cov}\,(\boldsymbol{\varepsilon}^*(t), \boldsymbol{\varepsilon}^*(t)) = \Sigma_{\boldsymbol{\varepsilon}}^{**}$. We have for given $\boldsymbol{\alpha}(t)$ (letting context determine whether a quantity such as $\mathbf{x}_t$ is random or observed),

$$\begin{pmatrix} \mathbf{a}_{t|t} \\ \mathbf{x}_t \\ \mathbf{x}_t^* \end{pmatrix} \sim N\left[\begin{pmatrix} I \\ H \\ H^* \end{pmatrix} \boldsymbol{\alpha}(t), \begin{pmatrix} C_{t|t} & 0 & 0 \\ 0 & \Sigma_{\boldsymbol{\varepsilon}} & \Sigma_{\boldsymbol{\varepsilon}}^* \\ 0 & \Sigma_{\boldsymbol{\varepsilon}}^{*T} & \Sigma_{\boldsymbol{\varepsilon}}^{**} \end{pmatrix}\right] \tag{6.1}$$

The conditional expectation of $\mathbf{x}_t^*$ is

$$E[\mathbf{x}_t^* \mid \mathbf{a}_{t|t}, \mathbf{x}_t] = E[\mathbf{x}_t^* \mid \mathbf{x}_t] = H^*\boldsymbol{\alpha}(t) + \Sigma_{\boldsymbol{\varepsilon}}^{*T}\Sigma_{\boldsymbol{\varepsilon}}{}^{-1}(\mathbf{x}_t - H\boldsymbol{\alpha}(t))\,. \tag{6.2}$$

On replacing $\boldsymbol{\alpha}(t)$ by $\mathbf{a}_{t|t}$ and substituting the MLE's of the covariogram parameters into $\Sigma_{\boldsymbol{\varepsilon}}$ and $\Sigma_{\boldsymbol{\varepsilon}}^*$ (signaled with a $\widehat{\ }$), the *Kriging predictor*, denoted $\widehat{\mathbf{x}}_{t|t}^*$, is

$$\widehat{\mathbf{x}}_{t|t}^* = H^*\mathbf{a}_{t|t} + \widehat{\Sigma}_{\boldsymbol{\varepsilon}}^{*T}\widehat{\Sigma}_{\boldsymbol{\varepsilon}}{}^{-1}(\mathbf{x}_t - H\mathbf{a}_{t|t})\,. \tag{6.3}$$

The estimated *Kriging variance* is

$$(\widehat{\Sigma}_{\boldsymbol{\varepsilon}}^{**} - \widehat{\Sigma}_{\boldsymbol{\varepsilon}}^{*T}\widehat{\Sigma}_{\boldsymbol{\varepsilon}}{}^{-1}\widehat{\Sigma}_{\boldsymbol{\varepsilon}}^*) + (H^* - \widehat{\Sigma}_{\boldsymbol{\varepsilon}}^{*T}\widehat{\Sigma}_{\boldsymbol{\varepsilon}}{}^{-1}H)C_{t|t}(H^* - \widehat{\Sigma}_{\boldsymbol{\varepsilon}}^{*T}\widehat{\Sigma}_{\boldsymbol{\varepsilon}}{}^{-1}H)^T. \tag{6.4}$$

6.2 Structure of the spatio-temporal 'covariance' function

From $\boldsymbol{\alpha}(1) = P\boldsymbol{\alpha}(0) + \boldsymbol{\eta}$ we have $\mathrm{var}\,(\boldsymbol{\alpha}(1)) = P\Sigma_0 P^T + \Sigma_{\boldsymbol{\eta}}$, and from $\boldsymbol{\alpha}(2) = P\boldsymbol{\alpha}(1) + \boldsymbol{\eta}$ we have $\mathrm{var}\,(\boldsymbol{\alpha}(2)) = P\mathrm{var}\,(\boldsymbol{\alpha}(1))P^T + \Sigma_{\boldsymbol{\eta}} = P^2\Sigma_0(P^2)^T +$

$P\Sigma_{\boldsymbol{\eta}}P^T + \Sigma_{\boldsymbol{\eta}}$. Define $Q_t = \mathrm{var}\,(\boldsymbol{\alpha}(t))$. Inductively,

$$Q_t = \mathrm{var}\,(\boldsymbol{\alpha}(t)) = P^t\Sigma_0(P^t)^T + \sum_{k=1}^{t} P^{k-1}\Sigma_{\boldsymbol{\eta}}(P^{(k-1)})^T\,. \tag{6.5}$$

Then, from (6.5) and the observation equation (2.2)–(2.3), for monitored and unmonitored locations, at the same time t, $t \le T$,

$$\begin{aligned} \mathrm{cov}\,(\mathbf{x}_t, \mathbf{x}_t) &= HQ_tH^T + \Sigma_{\boldsymbol{\varepsilon}} && (6.6)\\ \mathrm{cov}\,(\mathbf{x}_t, \mathbf{x}_t^*) &= HQ_tH^{*T} + \Sigma_{\boldsymbol{\varepsilon}}^* && (6.7)\\ \mathrm{cov}\,(\mathbf{x}_t^*, \mathbf{x}_t^*) &= H^*Q_tH^{*T} + \Sigma_{\boldsymbol{\varepsilon}}^{**}\,. && (6.8) \end{aligned}$$

For monitored sites at time t, $t \le T$, and unmonitored (say) sites at time $t+k$ $(k > 0)$ the covariance is

$$\mathrm{cov}\,(\mathbf{x}_t, \mathbf{x}_{t+k}^*) = HQ_t(P^k)^TH^{*T}\,. \tag{6.9}$$

We now consider a particular case, that of the autoregressive model, (4.6), where each common field follows a separate AR(1) model, and thus $P = \mathrm{diag}\,\{\phi_1, \ldots, \phi_{q+r}\}$. Further we take $\Sigma_{\boldsymbol{\eta}} = \sigma_\eta^2 I_p$, say, and $K = I_p$, where I_p is the $p \times p$ identity matrix. Suppose $|\phi_j| < 1$ so that the model is stationary, and the term $P^t\Sigma_0P^{tT}$ is negligible for large t, and

$$Q_t = \sum_{j=1}^{t} \mathrm{diag}\,\{\phi_1^{2j-1}, \ldots, \phi_p^{2j-1}\} = \mathrm{diag}\,\{b(\phi_1), ..., b(\phi_p)\}\,, \tag{6.10}$$

where $b(\phi_j) = \sum_{i=1}^{t} \phi_j^{2i-1} = \phi_j(1-\phi_j^{2t})/(1-\phi_j^2)$, $j = 1, \ldots, p$. This gives simplified forms of (6.6)–(6.9), for example, letting $\mathbf{h}_j^+$ and $\mathbf{h}_j^{*+}$ denote the j'th columns of H and H^* respectively,

$$\mathrm{cov}\,(\mathbf{x}_t, \mathbf{x}_t^*) = \sum_{i=1}^{p} b(\phi_j)\mathbf{h}_j^+\mathbf{h}_j^{*+} + \Sigma_{\boldsymbol{\varepsilon}}^* \tag{6.11}$$

Let $\mathbf{s}$ and $\mathbf{s}'$ be any two spatial locations, and write

$$\tau_t(\mathbf{s}, \mathbf{s}') = \mathrm{cov}\,(x(\mathbf{s}, t), x(\mathbf{s}', t)).$$

To simplify the exposition, suppose that the common fields are precisely the $p = m$ principal fields, (3.10). Then, from Section 3,

$$\tau_t(\mathbf{s}, \mathbf{s}') = \boldsymbol{\sigma}_\zeta(\mathbf{s})UQ_tU^T\boldsymbol{\sigma}(\mathbf{s}') + \sigma_\varepsilon(\mathbf{s}, \mathbf{s}')\,. \tag{6.12}$$

The first term of this covariance is seen to be a weighted mixture of products of the covariances between $\mathbf{s}$ and $\mathbf{s}'$ and each of the m normative sites. In particular for $\phi_i = \phi$, $i = 1, \ldots, p$, the covariance becomes

$$\tau_t(\mathbf{s}, \mathbf{s}') = b(\phi) \sum_{i=1}^{p} \sigma_\zeta(\mathbf{s}_i^*, \mathbf{s}) \sigma_\zeta(\mathbf{s}_i^*, \mathbf{s}') + \sigma_\varepsilon(\mathbf{s}, \mathbf{s}') \,, \tag{6.13}$$

again, for a stationary process. We can deduce some order relationships across time for τ_t (or for the corresponding correlations). For example, suppose that the summation in (6.13) is positive. Then, with $\phi > 0$, $\tau_t(\mathbf{s}, \mathbf{s}')$ is monotonically increasing with time, and $b(\phi)$ tends to the limit $\phi/(1-\phi^2)$.

7 Example

The data for this example comprise $T = 707$ daily observations of sulphur dioxide levels at $n = 14$ monitored sites in Leeds and its vicinity. These are almost two years of readings for the period 29th March 1983 to 5th March 1985. We first discuss selecting appropriate trend fields (Section 7.1), an appropriate covariogram model (Section 7.2) and the appropriate choice of number and nature of the principal fields (Section 7.3). For the chosen setup, we use Kalman filter recursions as part of maximum-likelihood estimation using the EM algorithm. We present various diagnostic displays, including plots of the predicted surface across several time points.

The positions of the 14 sites are shown in Figure 1. We see that 12 of the sites fall into two clusters, one cluster of 7 sites covering the city of Leeds itself (sites 6-12), the other cluster to the south covering part of the mining area to the south east of Leeds (sites 1-5).

Figure 2 shows the log-transformed values for the fourteen time series. There are several missing values indicated by gaps in the series. The sites Leeds 26 (row 3 column 3 in Figure 2) and Leeds 30 (row 4 column 2) are examples of series with large gaps. The pollution level is higher in the first year than in the second for each of the sites. The levels in the second summer are notably lower than in the first summer or during each winter.

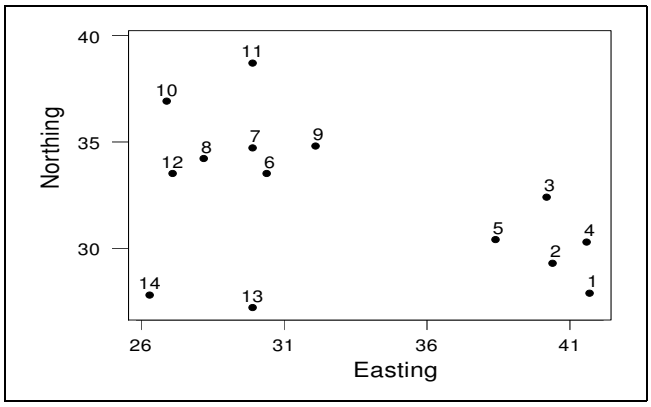

Figure 1: *Map of the sites where the data was recorded. The names of the sites are:- (1) Allerton Bywater 1, (2) Allerton Bywater 2, (3) Garforth, (4) Kippax, (5) Swillington,(6) Leeds 4, (7) Leeds 18, (8) Leeds 24, (9) Leeds 26, (10) Leeds 28, (11) Leeds 30, (12) Leeds 34, (13) Leeds 33, and (14) Morley. The axis are the labelled according to the grid references of the sites based on the Ordinance Survey grid that covers the United Kingdom*

7.1 Choice of trend fields

As a first step, prior to modelling spatial covariance, we considered alternative sets of trend fields, including constant and linear trend fields. For the example we chose to use a single, constant field. A couple of comments will help make sense of this choice. Firstly, a constant field reflects overall level in the series. Were we to model the average daily sulphur dioxide reading at the 14 sites as an autoregressive time series, then the terms of a state space forumulation of this model would plug directly into the specification (4.1)–(4.2), including $h_1(\mathbf{s}) \equiv 1$ and some number of zero fields. (In fact we do not use this to build the model, but it is easy to see that this strategy could continue, with a model for the next common field, etc.)

Our second comment is that the amplitude of each of the principal fields tends to dampen away from the neighborhood of each of the normative sites. This is implied by the form $\boldsymbol{\sigma}_\zeta(\mathbf{s})^T \mathbf{u}_{i_k}$, (3.10), of the principal field, and is evident in Figure 5 (below). A linear, quadratic, or other polynomial trend does not have this property. The relationship of Kriging solutions to spline fits is discussed by Kent and Mardia (1994) and Mardia et al (1996).

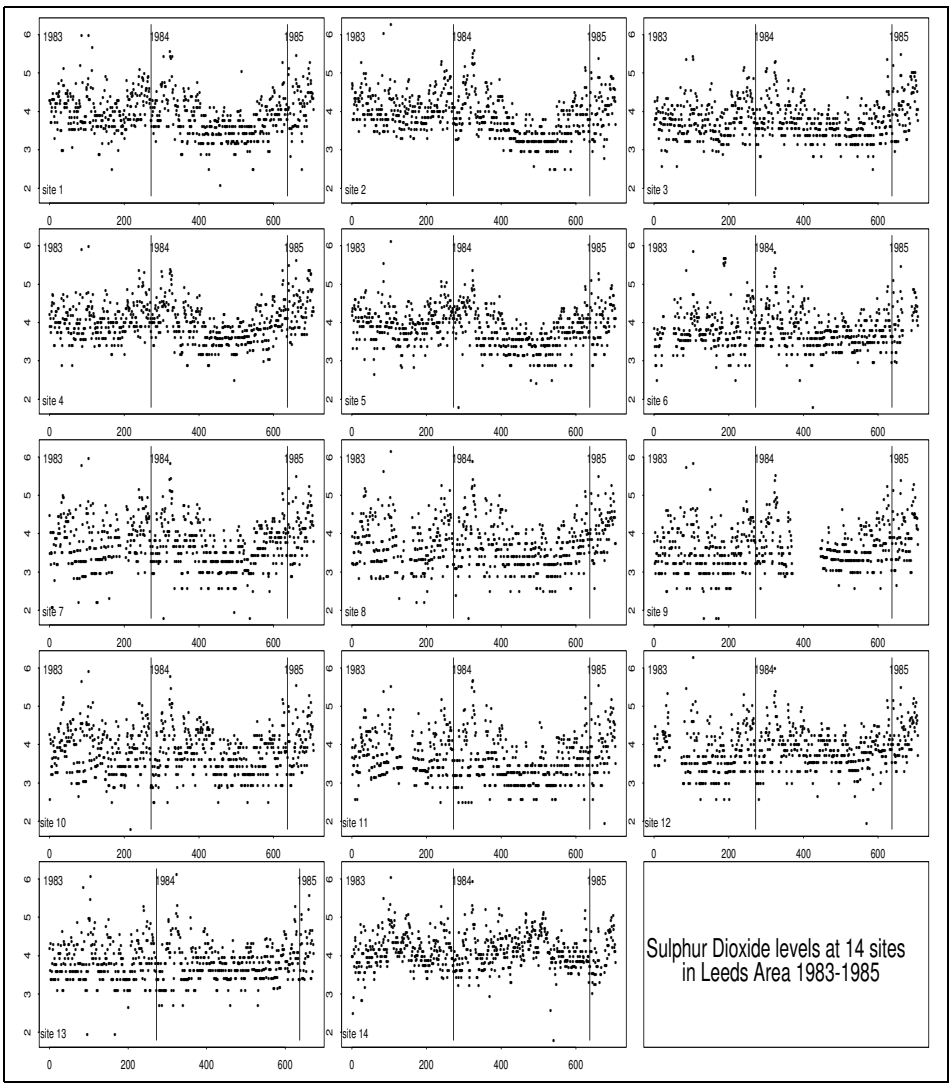

Figure 2: *Log transformed Sulphur Dioxide levels for fourteen sites in the neighbourhood of Leeds for the period March 29 1983 - March 5 1985. Sites are in order starting top left and reading across:- (1) Allerton Bywater 1, (2) Allerton Bywater 2, (3) Garforth, (4) Kippax, (5) Swillington,(6) Leeds 4, (7) Leeds 18, (8) Leeds 24, (9) Leeds 26, (10) Leeds 28, (11) Leeds 30, (12) Leeds 34, (13) Leeds 33, and (14) Morley.*

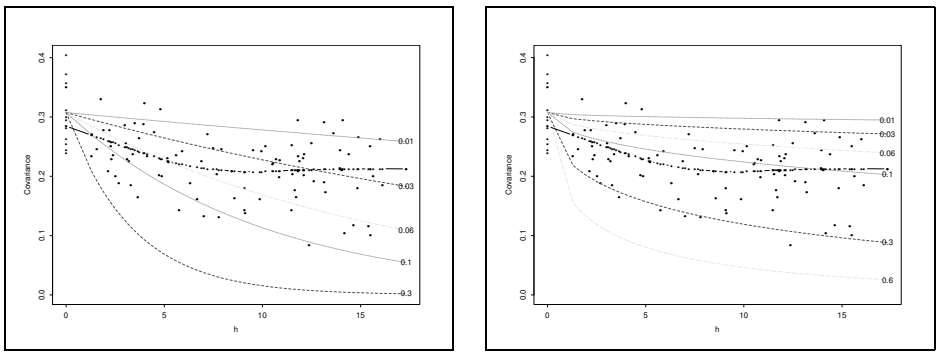

Figure 3: Covariogram plot with (1) $e^{-b|h|}$ and (2) $e^{-b\sqrt{h}}$ for different values of b. The values of b are indicated at the end of each curve. The more solid line is the loess curve.

As an aside, note that the principal fields comprise a basis for all possible sets of readings at the 14 sites orthogonal to the trend fields. These sets include any linear trend across the 14 sites, when linear trend is not one of the trend fields, but a linear trend across the 14 sites dampens to the overall average at spatial locations away from any of the 14 sites, and does not imply a global linear trend.

7.2 Choice of covariogram model

To model spatial covariance we use the exponential powered covariogram from the stable family (see Diggle *et al.*, 1998)

$$e^{-b|h|^p}, \quad b > 0, \quad 0 < p \leq 2\,. \tag{7.1}$$

This covariogram was compared, for various values of b and p (Figure 3), to the empirical covariogram obtained for the data aggregated into chunks each comprising about 50 time points. The empirical covariogram was more compact for log transformed data, so this was used. The model with $p = \frac{1}{2}$ gives the better fit.

Figure 3 superimposes the empirical covariogram and the theoretical covariograms for $e^{-b|h|}$ $(p = 1)$ and $e^{-b\sqrt{h}}$ $(p = \frac{1}{2})$. The smoothed estimate is also shown, the same in both plots. The closest fit is $p = \frac{1}{2}$ and $b = 0.1$.

Subsequently we found that the MLE with $p = \frac{1}{2}$ is $b = 1.3$. The difference between the preliminary ($b = 0.1$) and final ($b = 1.3$) models is explained by the use of the aggregated data over the full period. An exploration using subsets of the data suggests that, based on different periods of aggregation the shape of the variogram varies considerably. The length of the period of aggregation may result in over-estimating of the degree of association between neighbouring sites.

7.3 Choice of principal fields

The next stage is to choose the principal fields, given the particular choices of trend fields – the constant trend only – and the specific pair of values for p and $b - p = \frac{1}{2}$ and $b = 1.3$. These choices have taken us through rows 1-3 of Table 1. The next steps are to choose the number (m) and locations ($\mathbf{s}_i^*$) of the normative sites. Our initial approach, which we present here, is to use the $n = 14$ sites as $m = 14$ normative sites. The next step, row 6 of Table 1, is to select the principal fields. One of the principal fields will have zero eigenvalue, and is a constant corresponding to the constant trend field.

One possible criterion is to rank the principal fields in decreasing order of the variances of their coefficients when the observations at each time ($\mathbf{x}_t$) are expressed as a linear combination of the principal fields. But these variances are the eigenvalues of the covariogram, *i.e.* the reciprocal of the eigenvalues of B, the generalized bending energy matrix. A high degree of variability indicates a more global contrast over the set of sites while a small degree of variability suggests that the principal field is based on local comparisons. In other words, the principal fields corresponding to the largest eigenvalues of the generalized bending energy matrix result in local contrasts while those for the smallest non-zero eigenvalues capture the broader contrasts across the sites. Indeed Bookstein (1989) notes for the thin plate splines the association of the large eigenvalues with local variability and the small eigenvalues with global features.

The scree plot of principal field variance against rank of the non-zero eigenvalue, Figure 4, for the model $e^{-b\sqrt{h}}$ with $b = 1.3$, suggests that we include at least the first three or four principal fields. We use a total of five common fields, the constant trend surface and four principal fields.

The four principal fields corresponding to the smallest non-zero eigen-

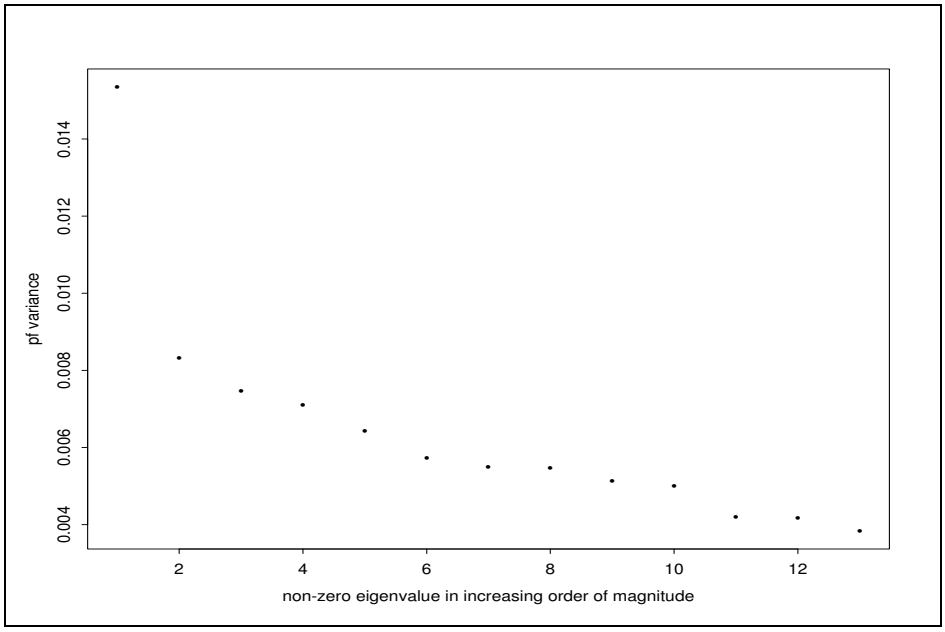

Figure 4: Plot of the variance in the coefficients in the principal fields against the order of the non-zero eigenvalues in increasing order of magnitude for the covariogram model $exp(-b\sqrt{h})$ *for* $b = 1.3$.

values of the generalized bending energy matrix B are shown in Figure 5. The value $b = 1.3$ is the MLE for the model with constant trend and the four principal fields with smallest non-zero eigenvalues. The principal field corresponding to the smallest non-zero eigenvalue is a north-south contrast between the group of 7 sites (6-12) centered on the City of Leeds and the group of 5 sites (1-5) towards the south-east. This area was the principal mining area near to Leeds. The next principal field is a contrast between the two groups totaling 12 sites and the remaining two sites (13 and 14) in the south-west. The third and fourth principal fields are local north-south and east-west contrasts within the seven sites (6-12) that cover the city of Leeds.

The values of the five common fields at the 14 sites are tabulated in Table 2, and are based on (3.10) of Section 3. The values also appear as "+" symbols in Figure 5.

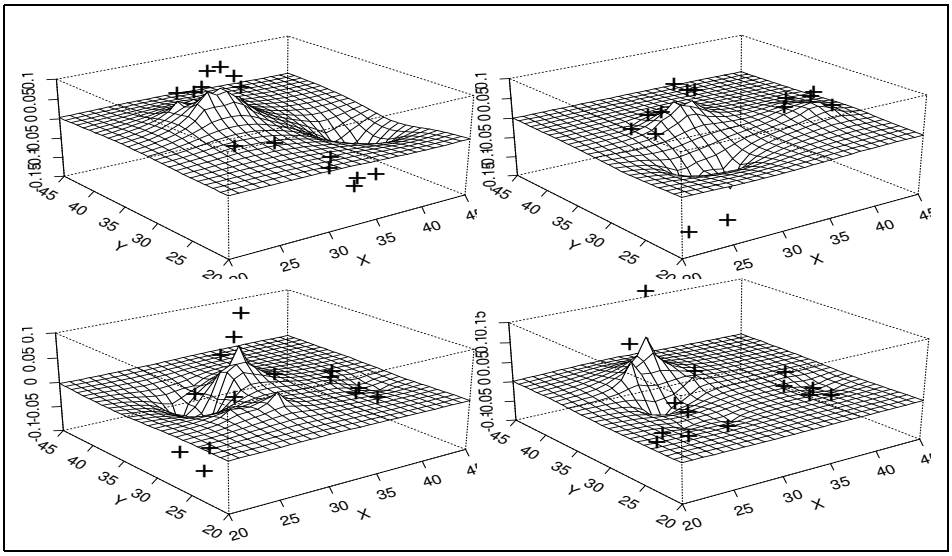

Figure 5: *The four principal fields with the smallest non-zero eigenvalues of the bending energy matrix based on the model* $e^{-b\sqrt{h}}$ *with* $b = 1.3$ *after removal of a constant spatial trend.*

	site name	$h_1(\mathbf{s})$	$h_2(\mathbf{s})$	$h_3(\mathbf{s})$	$h_4(\mathbf{s})$	$h_5(\mathbf{s})$
1	Allerton Bywater 1	1.0	-0.140	0.039	-0.024	-0.029
2	Allerton Bywater 2	1.0	-0.174	0.059	-0.019	-0.032
3	Garforth	1.0	-0.120	0.034	0.009	0.006
4	Kippax	1.0	-0.166	0.571	-0.020	-0.026
5	Swillington	1.0	-0.123	0.334	0.015	-0.005
6	Leeds 4	1.0	0.130	0.942	0.110	-0.050
7	Leeds 18	1.0	0.146	0.103	0.071	-0.034
8	Leeds 24	1.0	0.147	0.047	-0.108	-0.098
9	Leeds 26	1.0	0.087	0.079	0.144	0.037
10	Leeds 28	1.0	0.084	-0.008	-0.127	0.116
11	Leeds 30	1.0	0.060	-0.000	-0.030	0.218
12	Leeds 34	1.0	0.121	0.004	-0.147	-0.108
13	Leeds 33	1.0	0.012	-0.183	0.075	-0.035
14	Morley	1.0	0.021	-0.195	0.043	-0.042

Table 2: *Values of the principal fields (the matrix* H*)*

7.4 Parameter estimation and model interpretation

With the five common fields identified above, and the matrix H shown in Table 2, we turn to the temporal modelling components of KKF. The initial model specification step, row 7 of Table 1, can lead to a variety of alternatives. In particular a degree of temporal memory could be introduced (Section 4.1) using zero fields and additional ($p > 5$) state parameters to code autoregressive models for any or all of the original 5 common fields.

We consider the case when the transition matrix P is unconstrained and 5×5. The innovation covariance matrix $\Sigma_{\boldsymbol{\eta}}$ is constrained only to be non-negative definite. Further, $\Sigma_{\boldsymbol{\varepsilon}}$ is assumed to be diagonal, meaning that all the spatial structure is encapsulated in the principal fields. With these provisions, the temporal parameters $\boldsymbol{\theta}$ comprise P, $\Sigma_{\boldsymbol{\eta}}$, and the diagonal elements of $\Sigma_{\boldsymbol{\varepsilon}}$.

We use the EM algorithm and the Kalman filter recursions to maximize the likelihood over $\boldsymbol{\theta}$ for several values of the spatial covariogram parameter b. The minimum value of $-2 \ln L$, (5.13), is -13620.28, and, at this value, $b = 1.3$. Figure 6 plots $-2 \ln L$ against b.

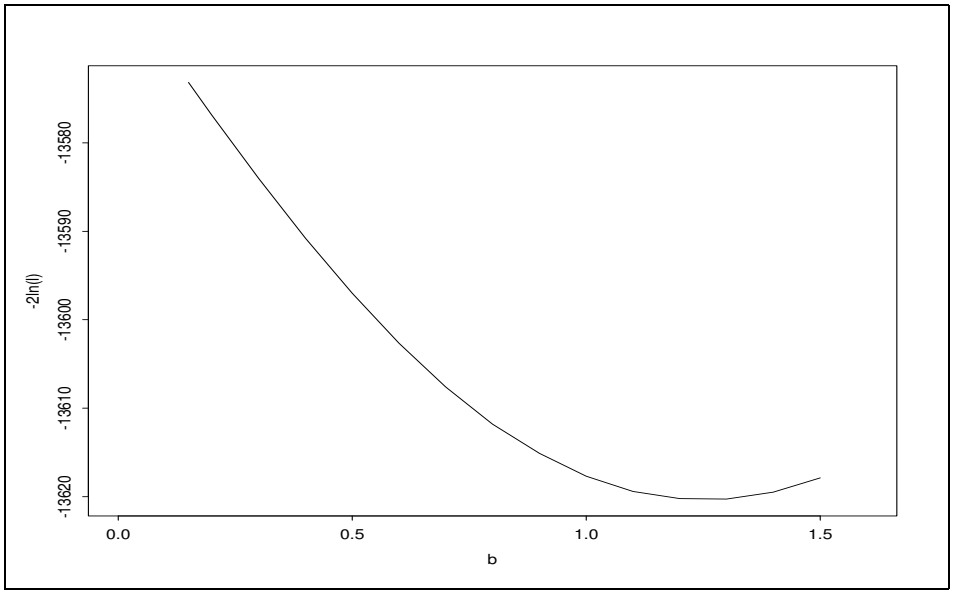

Figure 6: $-2ln(\ell)$ *versus* b *for the model* $e^{-b\sqrt{h}}$

The three parts of the maximum-likelihood estimates of $\boldsymbol{\theta} = (P, \Sigma_{\boldsymbol{\eta}}, \Sigma_{\boldsymbol{\varepsilon}})$ are as follows. The estimated transition matrix is

$$\widehat{P} = \begin{bmatrix} 0.6467 & -0.0715 & -0.0035 & -0.1393 & -0.0791 \\ 0.0162 & 0.7727 & -0.0493 & -0.0349 & -0.1707 \\ -0.0823 & -0.0659 & 0.9301 & -0.1558 & -0.0952 \\ -0.1048 & 0.0998 & -0.0157 & 0.8720 & 0.0942 \\ 0.0451 & -0.0934 & -0.0342 & -0.0386 & 0.8104 \end{bmatrix}. \tag{7.2}$$

The estimated innovation covariance matrix is

$$\widehat{\Sigma}_{\boldsymbol{\eta}} = \begin{bmatrix} 0.1216 & & & & \\ 0.0690 & 0.2456 & & & \\ 0.0830 & 0.0176 & 0.1438 & & \\ 0.0126 & -0.1398 & 0.0660 & 0.1528 & \\ 0.4610 & 0.1011 & 0.0468 & -0.0596 & 0.0964 \end{bmatrix}. \tag{7.3}$$

The estimated diagonal terms of the measurement error matrix, $(\widehat{\Sigma}_{\boldsymbol{\varepsilon}})_{ii}$, $i = 1, \ldots, 14$, are 0.052, 0.055, 0.065, 0.032, 0.061, 0.099, 0.074, 0.056, 0.071, 0.099, 0.046, 0.047, 0.064, and 0.085.

In writing down the covariogram for the principal fields, Σ_{ζ}, we assumed that the time series at each site have approximately equal variances. In fact, the covariogram plot, Figure 3 shows this to be the case, with a less-than-two-fold difference in variances. After removing the mean component, (3.8), of the KKF model, the error variances are similarly approximately equal. The range in the maximum-likelihood estimates is three-fold, 0.032-0.099.

The time series of estimated states $\widehat{\alpha}_j(t) = (\mathbf{a}_{t|t})_j$, $j = 1, \ldots, 5$ and $t = 1, \ldots, 707$ are shown in Figure 7. The diagonal elements predominate in the transition matrix, and the innovation variances are small relative to the state estimates.

A picture of the features in the data begins to emerge. The first panel (row 1 col 1), for the constant trend field, shows that there is considerable temporal variation in the overall level, with occasional high spikes, and fewer departures where the overall level is below average. The highest values are scattered throughout the first 12 months. There is also an upward trend in the winter 1984/85. For the first principal field, the north-south contrast, shown in the second panel (row 1 col 2), there is less variability in the weights in the second part of the time period suggesting that this contrast had considerably more "random" effect in the first year of the data

values compared with the second. There is a systematic, though decreasing, positive effect in the second year. For the second principal field, the east-west contrast shown in the third panel (row 2 col 1) we see that the effect in the two winter periods is the reverse of its contribution in the summer periods. This may relate to different patterns of prevailing winds. For the third and fourth principal fields any pattern is harder to see. This may be because the contrasts are over a relatively smaller distance.

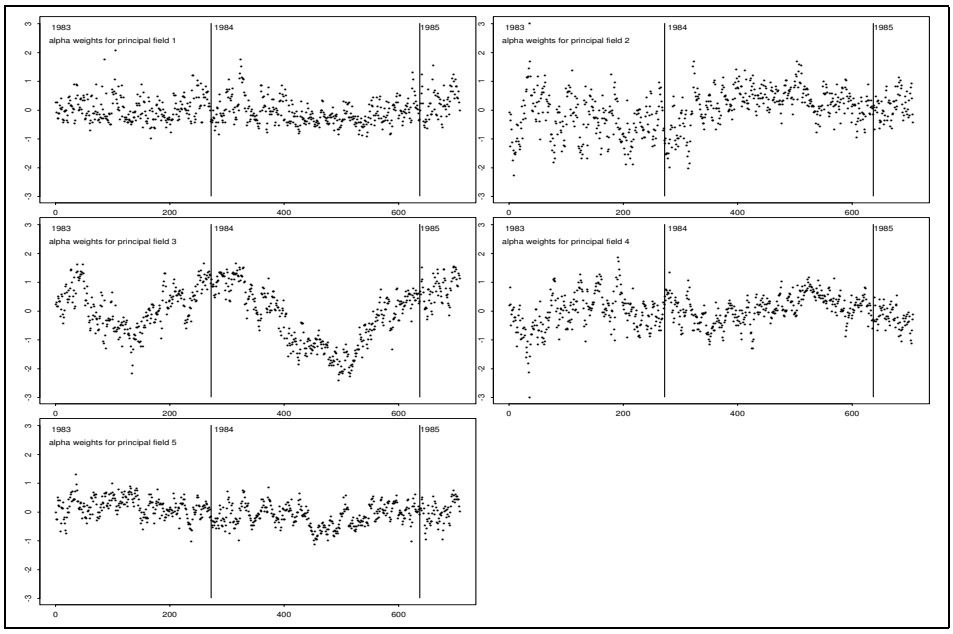

Figure 7: Estimated values of the states $\mathbf{a}_{t|t}$

The residuals from the fitted model are shown in Figure 8. These suggest that the model gives a similar accuracy of fit to all the time series. The residual sum of squares (RSS) for each series, shown in Table 3, ranges from 12.3 to 42.3, with an average of 25.35.

7.5 Surface interpolation and prediction

Figure 9 shows the reconstructed surface over a 9 day period starting at the 318th day. The selection of $t = 318$ as starting point was arbitrary.

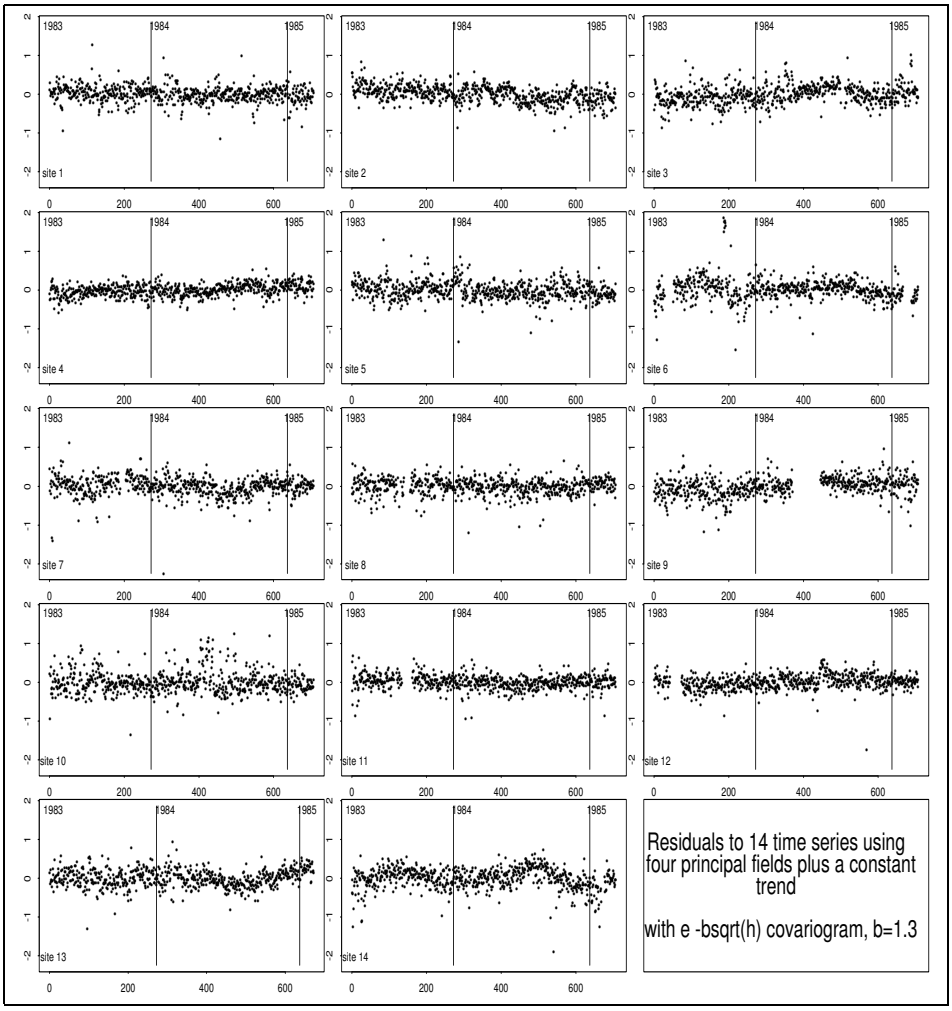

Figure 8: *Residual values for the fourteen sulphur dioxide series based on 5 principal fields determined using the model* $e^{-b\sqrt{h}}$ *with* $b = 1.3$

site	1	2	3	4	5	6	7
RSS	20.06	22.46	27.19	12.29	27.84	39.01	32.16
site	8	9	10	11	12	13	14
RSS	17.51	21.92	34.45	13.37	18.29	26.10	42.31

Table 3: *Residual sums of squares*

The values displayed are

$$\widehat{x}(\mathbf{s},t) = \sum_{j=1}^{5} h_j(\mathbf{s})\widehat{\alpha}_j(t)\,, \tag{7.4}$$

as $\mathbf{s}$ ranges over a 25×25 grid that covers the region.

Errors are available provided $\Sigma_{\boldsymbol{\varepsilon}}$ is interpolated to a covariogram, possibly by assuming uncorrelated isotropic measurement error at every spatial location.

Figure 9 also gives the pollution levels relative to the general surface mean below each surface plot. It can be seen that during the 9 day period the levels were near average for the first few days and then rose to a peak on days 324 and 325. When the pollution levels were highest the peak sulphur dioxide concentration was over the city, while at lower pollution levels the peak shifted to the mining region.

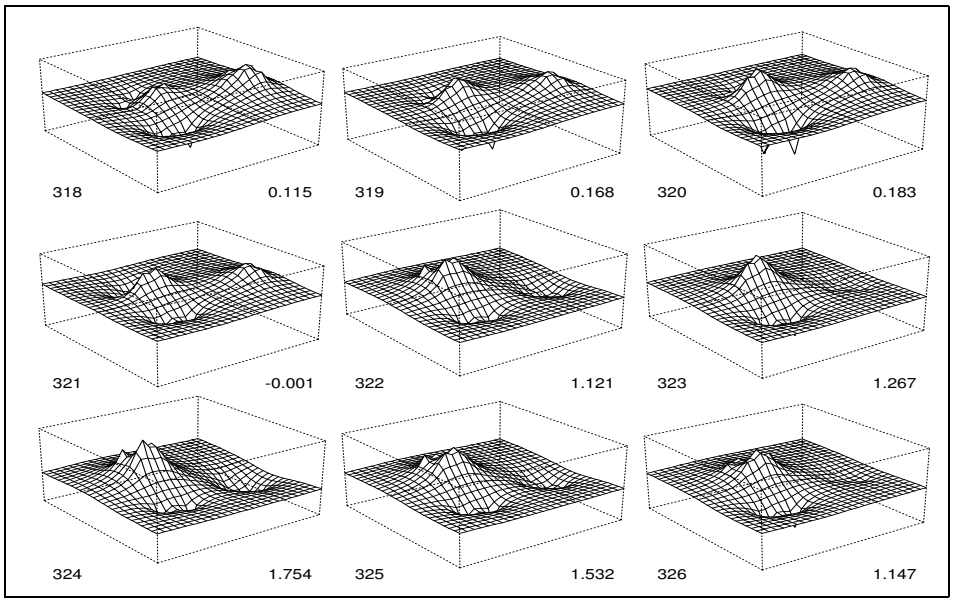

Figure 9: Fitted surface for sulphur dioxide levels based on 5 principal fields determined using the model $e^{-b\sqrt{h}}$ - the two values on each graph indicate the time point and the distance of the mean of the surface from the mean of the data

8 Discussion

We have pursued spatial-temporal modelling using KKF from start to finish, both theoretically and by way of an example. We have developed some properties of KKF. However, we have just scratched the surface of this complicated situation. Also, we have worked here on a single practical example but in future we hope to implement the model on a range of practical examples. The area of model selection in particular is considerable, combining the possibilities of spatial prediction (Kriging) with temporal modelling. The multivariate aspect of the temporal model follows from the use of a set containing more than one common field, even in our setting which has the observation at each site to be univariate. Multivariate time series modelling opens up the possibility of a different model for each common field, with selective cross terms.

We have used the powered exponential covariance function (7.1) in our work. This covariance function is continuous for all p, but it is not differentiable for $p < 2$. For $p = 2$ the corresponding covariance matrix becomes grossly ill-conditioned (Kent, 1998). Also, in practice, there can be confounding between the choice of the 'range' parameter b and the 'smoothing' parameter p. A more flexible model we could have used is the Matern covariance function. Anisotropic models are also relevant in general, but this question is not so important when using only a small number of normative sites for principal fields. Details of diagnostic tools need to be explored in future for covariance selection (cf. Handcock and Stein, 1994).

The choice of common fields is another rich area to be explored. The combination of trend fields and principal fields provides overall a statistically well-founded approach. In particular cases a more specific choice of fields may be called for. We discuss the dynamic linear model (Section 4.2) in this context. A related approach is, in modelling seasonal data, to suppose that there are two or more common fields each specific to one season; the general state space model governs the rise and decline of the weights attached to each of these fields through time, possibly taking the form, for monthly data, of an ARMA model with periodicity 12.

Goodall and Mardia (1994) discuss other other approaches to determining common fields. One uses empirical orthogonal functions. This approach has been implemented by Wikle and Cressie (1997). Another approach, that of Cox and Isham (1988), is to use spatial kernel functions to define

the common fields.

Goodall and Mardia (1994) provide various extensions that have not been pursued here. For example modelling multivariate data, the extension is conceptually straightforward but as expected there are implementation problems. For extensions to cases where the observations include images, "images as data", there are more complications when space S is to be linked with principal fields. Are the blocks taken over the whole pixel? In general, we can use block Kriging for principal fields.

We have used the linear Kalman filter but it can be extended in various directions, for example to the non-linear Kalman filter on the line of Durbin and Koopman (1997). Also we can allow non-Gaussian spatial observations following the model of Diggle *et al.* (1998).

The optimal property of the Kriging predictor, and the optimality of the Kalman filter, are well known separately. It is expected that their properties will be carried over into the combined KKF setting. However, since we are only using selected principal fields, rather than the complete spatial information, we expect some additional requirements for optimality. The predictor and its error will, of course, be influenced by parameter estimation, but the predictors errors will tend to be underestimated.

Our model parameters can be estimated using MCMC techniques adapting the ideas of Diggle *et al.* (1998) and Green (1995). However, the problem of model selection requires further modification. Recent Bayesian methods might provide an alternative strategy (see, for example, Gelfand *et al.*, 1992). To obtain the parameters in the covariance functions using maximum-likelihood estimation requires in general inverting the site covariance matrix which can have very high dimension. This creates numerical problems (see, Mardia and Marshall, 1984). However our method of using principal fields circumvents this difficulty.

We have given some thought to the selection of normative sites for principal fields, including both sites included in the original design (design sites), and non-design sites. Further work is required. Suppose m normative sites are to be selected from among the n design sites, and a total of $p = m$ common fields used, to include a fixed set of q trend fields, and $r = m - q$ principal fields. Consider all $\binom{n}{m}$ subsets written $\mathbf{i} = (i_1, \ldots, i_m)$ of the sites indexed $(1, \ldots, n)$. For each $\mathbf{i}$ compute the generalized bending energy matrix $B_{\mathbf{i}}$ (3.5) and its largest eigenvalue, written $d_{m\mathbf{i}}$. The selected

sites are $\mathbf{i} = \arg\min_{\mathbf{i}} d_{m\mathbf{i}}$, a minimax solution. This selection seems to concentrate sites on boundary of convex hull which may not be appropriate. Diagnostics along the lines of Section 4.2 of Goodall and Mardia (1994) should be considered.

Acknowledgements

Kanti Mardia would like to thank José M.Angulo for his encouragement in writing this paper, and to thank Ramón Gutiérrez-Jáimez for his support to visit Universidad Granada. The work is also supported by travel funds of Leeds University, Department of Statistics, and by National Science Foundation award DMS92-08656 to the Pennsylvania State University, Department of Statistics, and by a short-term visiting position awarded to Colin Goodall to Bristol University, Department of Mathematics. F.J. Alonso was also supported by project PB96-1440 of the Dirección General de Enseñanza Superior, CICYT, Ministerio de Educación y Cultura, Spain

References

Akaike, H. (1973). Information theory and an extension to the maximum likelihood principal. *Second International Symposium on Information Theory* (B.N.Petrov and F. Csaki, editors). Akademiai Kiado, Budapest, 267–281.

Ansley, C. F. and R. Kohn (1984). On the estimation of ARIMA models with missing values. *Time series analysis of irregularly spaced data* (E. Parzen, editor). Springer–Verlag, New York, 9–37.

Bookstein, F. L. (1989). Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 567-585.

Cohen, A. and R. H. Jones (1969). Regression on a random field. *Journal of American Statistical Association*, **64**, 1172-1182.

Cox, D. R. and V.S. Isham (1988). A simple spatial-temporal model of rainfall. *Proceedings of the Royal Society of London*, *A*, **415**, 317-328.

Cressie, N. A. C. (1993). *Statistics for Spatial data*, 2nd edition. John Wiley.

Creutin, J. D. and C. Oblen (1982). Objective analysis and mapping techniques for rainfall fields: An objective comparison. *Water Resources Research*, **18**, 413-431.

Dempster, A. P., N.M. Laird and D.B. Rubin (1977). Maximum Likelihood from

incomplete data via the EM algorithm. *Journal of the Royal Statistica Society, Series B*, **39**, 1–38.

Diggle, P.(1998). Space-time calibration of radar rainfall data. *Environmental Statistics Study Group talk, Royal Statistical Society.*

Diggle, P., J.A. Tawn, and R.A. Moyeed (1998). Model based geostatistics (with discussion). *Journal of the Royal Statistical Society, Series C*, **47**, 299-350.

Durbin, J. and S.J. Koopman (1997). Monte calrlo maximum likelihood estimation for non-Gaussian stats space models, *Biometrika*, **84**, 669–684.

Gelfand, A. E., D.K. Dey and H. Chang (1992). Model determination using predictive distributions with implementation via sampling based methods. *Bayesian Statistics, vol 4*, (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, editors), Oxford University Press, 147-167.

Goodall, C. and K.V. Mardia (1994). Challenges in Multivariate Spatial modelling. *Proceedings of the XVIIth International Biometric Conference*, Hamilton, Ontario, Canada, 8-12.

Green, P. J. (1995). Reversible jump Markov-chain Monte-Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Green, P. J. and D.M. Titterington (1988). Recursive methods in image processing. *Bulletin of the International Statistical Institute*, **52-4**, 51-67.

Gregoire, T. C., D.R. Brillinger, P.J. Diggle, E. Russek–Cohen, W.G. Warren and R.D. Wolfinger (1998). *Modelling Longitudinal and Spatially Correlated data : Methods Application and Future Directions*, Lecture Notes in Statistics, **122**, Springer Verlag, New York.

Gupta, N. K. and R.K. Mehra (1974). Computational aspects of maximum likelihood estimation and reduction in sensitivity calculations. *IEEE Transactions on Automatic Control*, **AC–19**, 774-783.

Guttorp, P. and P.D. Sampson (1994). Methods for estimating heterogeneous spatial covariance functions with enviromental applications. *Environmental Statistics* (G.P. Patil and C.R. Rao, editors). *Handbook of Statistics*, **12**, 661-689. North Holland, Amsterdam, New York.

Hancock, M.S. and M.L. Stein (1993). A Bayesian analysis of Kriging, *Technometrics*, **35**, 403-450.

Harrison, P. J. and C.F. Stevens (1976). Bayesian Forecasting, *Journal of the Royal Statistical Society, Series B*, **38**, 205–247.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, Cambridge.

Haslett, J. (1989). Space time modelling in meteorology - a review. *Bulletin of*

the International Statistical Institute, **51**, 229-246.

Janacek, G. and L. Swift (1993). *Time series: forecasting, simulations, applications*, Ellis Horwood, New York.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observation. *Technometrics*, **22**, 389–395.

Kent, J. T. (1998). Discussion to model-based Geostatistics by Diggle, P., Tawn, J.A. and Moyeed, R. A. *Journal of the Royal Statistical Society, Series C*, **47**, 330-331.

Kent, J. T. and K.V. Mardia (1994). The link between Kriging and thin-plate splines. *Probability, Statistics and Optimization* (F.P. Kelly, editor). John Wiley, New York, 324–329.

Mardia, K. V. and C.R. Goodall (1993). Spatial–temporal analysis of multivariate enviromental monitoring data. *Multivariate Enviromental Statistics* (G. R. Patil and C.R. Rao, editors). Elsevier.

Mardia, K. V., J.T. Kent, C.R. Goodall and J.A. Little (1996). Kriging and splines with derivative information, *Biometrika*, **83**, 207-221.

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

Meinhold, R. J. and N.D. Singpurwalla (1983). Understanding the Kalman filter. *American Statistician*, **37**, 123-127.

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

Rouhani, S. and D.E. Myers (1990). Problems in space-time Kriging of geohydrological data. *Math. Geology*, **22**,611-623.

Shumway, R.H. and D.J. Stoffer (1982). An approach to time series smoothing and forecasting using the EM algorithm, *Journal time series analysis*, **3**, 253–264.

Smith, R.L. (1998). Estimating nonstationary correlations. *Preprint.*

Watson, M. W. and R.F. Engle (1983). Alternative algorithms for the estimation of dynamic factor, MIMIC and varying coefficient regression. *Journal of Econometrics*, **23**, 385–400.

West, M. and P.J. Harrison (1989). *Bayesian Forecasting and Dynamic Linear Models*, Springer-Verlag, New York.

Wikle, C. K., L.M. Berliner and N.A.C. Cressie (1997). Hierarchical Bayesian time models. *Preprint.*

Wikle, C. K. and N.A.C. Cressie (1997). A Dimension-reduction approach to space-time Kalman filtering. *Preprint.*

José M. Angulo
Universidad de Granada, Spain

First of all, I want to congratulate the authors on this interesting paper, which I believe represents an important contribution to research on space-time processes. In the past few years, this area has received special attention in the literature, due to the interest in modeling and predicting variables representing phenomena evolving in both space and time in many fields of application, as pointed out by the authors. A recent interesting review on different approaches, such as those looking at time series correlated in space or at spatial processes evolving in time, and how different methodologies from time series analysis and from spatial statistics are integrated into space-time analysis, etc., can be found in Wikle and Cressie (1998).

The key point in the analysis of space-time data, and also its main source of complexity, as might be expected, consists in the appropriate definition and representation of spatio-temporal interaction, a non-trivial aspect which adds a new 'dimension' to the problem in contrast with purely spatial and/or temporal studies. As usual in many scientific works, it is necessary to establish a compromise between the complexity of the model and the computational efficiency. In Mardia *et al.*'s paper, the authors develop the 'Kriged Kalman filter' as a useful and general 'strategy' based on a combination of the two well-known Kriging and Kalman filter approaches, for space and time series analyzes, respectively.

The results in the paper derive from a long-term research wherein the authors have aimed at providing both generality and practical feasibility. I will confine my comments to four specific aspects that mainly concern the way the spatio-temporal interaction is modeled, the method implementation, and possible extensions.

First, both the assumption of a set of 'normative' sites on which to base the definition of the Kriging space for the spatial component, as well as the assumption of a fixed set of locations where observations are obtained over time —as the authors point out, they usually but do not necessarily coincide—, lead to adopting a multivariate consideration of time with respect to space. In contrast, the paper presents an essentially univariate

consideration of space with respect to time, in the large-scale variation, when assuming that there is a common spatial structure from which the common trend and principal fields are defined. Two interesting consequences derive from this setup: first, an elegant formulation where time and space are 'separated' in some way; second, some computational simplifications. In this respect, I think one can partially renounce the first aspect, without much trouble in computation except for computing time, in favour of a multivariate consideration of space with respect to time. Indeed, this may be more realistic for long time(-space) series, or just for series where the 'underlying' spatial dependence conditions may be subject to changes over time for different reasons. The question is, why not consider that the definition of the common fields, and not only their effect through the state component coefficient, can also adapt over time? As mentioned in the paper, Goodall and Mardia (1994) discuss this aspect, considering a general formulation consisting of an 'extended Kalman filter', which includes two sets of state equations related to $\boldsymbol{\alpha}(t)$ and $\mathbf{h}(\mathbf{s}, t)$. We can, however, consider adopting an intermediate solution, with a simpler treatment for the smooth time evolution of $\mathbf{h}(\mathbf{s}, t)$. For instance, fix the set of 'normative sites', the number and nature of the trend functions, and the number of principal fields, say corresponding to a fixed number of the smallest eigenvalues of the partial information matrix excluding the zero eigenvalues associated with the trend terms. One way to consider the 'common' spatial dependence conditions at each time would be to estimate the spatial covariances dynamically using all the data based on a moving time neighborhood (lag window). That is, matrix H will now be dependent on time, H_t —as considered in the usual space-state model formulation (*e.g.*, Harvey (1989))—, with H_t changing smoothly over time depending on the lag window considered. Depending on the degree of changes observed in H_t over time, and for the sake of saving computational time, we might decide to update H_t at intervals larger than unit, or when changes in coefficients measured in some way are larger than fixed tolerance values. Thus, the main structure of the model remains the same, but we introduce some dynamics in the 'common' fields ('common' would now have a local meaning) that may help to relax the imposition on $\boldsymbol{\alpha}$ and ε to represent the whole time evolution behaviour. The choice of a convenient lag window appears as a new selection problem. In particular, we may consider the use of seasonal lag windows to account for different spatial dependence patterns in different seasons, a situation which the authors refer to in the paper.

Related in part with my previous comment, one specific question on diagnostics arises concerning the example. In view of certain more or less slight patterns in some of the residual time series, we may at first interpret that the number of principal fields to be included in the mean component μ should be increased. A different interpretation, again, would be that there may exist different spatial patterns depending on the seasons and also on the years. Such a case could be handled under the dynamic approach to the common fields.

A second aspect I would like to comment on is the problem of selecting the set of normative sites on which the Kriging space is to be defined. As the authors acknowledge, further work is required in this respect. Consider the case of a unique 'static' bending energy matrix. Assuming the interpretation (Bookstein, 1989) that small eigenvalues of the bending energy matrix B are related with large-scale variation, while large eigenvalues are related with smaller-scale features, and assuming also that the mean component μ is meant to represent global behaviour, the question is: would it not be more reasonable to select the set $\mathbf{i}$ of normative sites such that, say, the sum of the p smaller eigenvalues is minimized (a different minimax criterion)? Or, assuming that there are no affine components in the residuals after removing trends, such that the product of the eigenvalues $d_{q+1}, ..., d_{q+r}$ is minimized (another minimax criterion)? This selection problem becomes, of course, more complex if we consider a dynamically adapted bending energy matrix, as I suggested above. A first solution can be given using the global bending energy matrix obtained from the whole data set, but it may be worthwhile to investigate strategies that take into account the dynamic approach.

One particular point that also attracted my attention in reading the paper was the attempt made in the example to give an interpretation of the selected principal fields in terms of spatial 'contrast' patterns. The definition of the principal fields as linear combinations of covariance functions with respect to fixed points (normative sites), with coefficients derived from the eigendecomposition of the matrix B, has both interesting technical properties and interpretations. In certain applications, however, it might be interesting to define such linear combinations using alternative decompositions for the bending energy matrix, possibly considering vectors of coefficients defined 'ad hoc' to represent local contrast, maybe allowing departure from orthogonality up to some extent. This is somehow related with the idea of using a wavelet decomposition. Such an approach could

be interesting, for instance, for comparing the results obtained for different spatio-temporal variables on the basis of a common set of normative sites and a common set of vectors for linear combinations. Further research along these lines may be of interest.

Finally, I would like to mention the possible use of covariates, when information is available, for improving the representation and estimation of the 'systematic behaviour' term μ, and hence of the model. This idea was introduced following a semiparametric approach in García-Jurado *et al.* (1995), in the time-series analysis context. An extension to the case of spatio-temporal data has recently been proposed in Angulo *et al.* (1998). We expect this added aspect to significantly improve the performance of the KKF approach in some situations, in both the estimation results and their interpretation.

In conclusion, I think this relevant work will clearly be of methodological interest to practitioners of spatio-temporal data analysis in many areas of application, in particular in environmental studies. There remain in and arise from this work a number of open problems concerning implementation and possible extensions that undoubtedly deserve attention for future research.

References

Angulo, J.M., W. González-Manteiga, M. Febrero-Bande and F.J. Alonso (1998). Semiparametric statistical approaches for space-time process prediction, *Environmental and Ecological Statistics*, **5**(4) (to appear).

Bookstein, F.L. (1989). Principal warps: Thin-plate splines and the decomposition of deformation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 567-585.

García-Jurado, I., W. González-Manteiga, J.M. Prada-Sánchez, M. Febrero-Bande and R. Cao (1995). Predicting using Box-Jenkins, nonparametric, and bootstrap techniques, *Technometrics*, **37**(3), 303-310.

Goodall, C.R., and K.V. Mardia (1994). Challenges in multivariate spatio-temporal modeling, *Proceedings of the XVIIth International Biometric Conference*, Hamilton, Ontario, Canada, 8-12 August 1994.

Harvey, A.C. (1989). *Forecasting, structural time series models, and the Kalman filter*, Cambridge University Press.

Wikle, C.K., and N. Cressie (1997). *A dimension-reduction approach to space-time Kalman filtering* (preprint).

N. Cressie
The Ohio State University, U.S.A.
C.K. Wikle
University of Missouri, U.S.A.

Strategies for dynamic space-time statistical modeling: discussion of "The Kriged Kalman filter" by Mardia et al.

We appreciate the opportunity to discuss this paper by Mardia, Goodall, Redfern, and Alonso (hereafter referred to as MGRA) because it allows us to put in perspective and, to some extent, prioritize the considerable recent activity in dynamic space-time prediction. MGRA represents one contribution, whose place among the recent contributions in the statistics literature by Cressie (1994), Goodall and Mardia (1994), Guttorp, Meiring, and Sampson (1994), Huang and Cressie (1996), Wikle and Cressie (1997), Berke (1998), Meiring, Guttorp and Sampson (1998), and Wikle, Berliner, and Cressie (1998) shall be discussed.

In this discussion we shall show that MGRA's contribution yields a space-time Kalman filter (STKF), which they call the Kriged Kalman filter, that in general oversmooths. We shall also emphasize the importance of incorporating the temporal dynamics into the space-time modeling and show how such dynamics imply the form of the STKF. All this is given within the context of a review of the STKF literature that has appeared in the recent past.

In a manuscript that is under revision for *Biometrika*, Wikle and Cressie (1997) present a space-time Kalman filter that is derived from a quite general dynamic space-time statistical model. We shall now rewrite the model of Wikle and Cressie (1997) in a notation that is comparable to MGRA's. There are three model equations,

$$x(\mathbf{s},t) = y(\mathbf{s},t) + \zeta(\mathbf{s},t) \tag{1}$$

$$y(\mathbf{s},t) = \mu(\mathbf{s},t) + \nu(\mathbf{s},t) \tag{2}$$

$$\mu(\mathbf{s},t) = \int \omega_{\mathbf{s}}(\mathbf{u})\mu(\mathbf{u},t-1)d\mathbf{u} + \eta(\mathbf{s},t), \tag{3}$$

where the error terms ζ, ν, and η are assumed Gaussian. In equation (1), the error term ζ is white noise and represents measurement error. In equation (2), the error term ν represents spatial structure that is independent from one time point to the next. In the terminology of Cressie (1993, Chapter 2), ν represents the small-scale spatial structure that, while exhibiting temporal variability, has no temporal dynamics associated with it. The space-time dynamics are found in the other component of (2), namely μ.

In equation (3), $\mu(\mathbf{s}, t)$ is seen to depend dynamically on $\mu(\mathbf{u}, t-1)$, where $\mathbf{u}$ belongs to some "neighborhood" of $\mathbf{s}$ determined by the support of the interaction function $\omega_{\mathbf{s}}(\mathbf{u})$. Thus, if rainfall is the variable of interest and weather patterns are generally from west to east, then the rainfall tomorrow at location $\mathbf{s}$ will depend on today's rainfall at locations $\mathbf{u}$ within a swath west of $\mathbf{s}$ up to a distance determined by the speed of the weather system. Importantly, the error term η in equation (3) is temporally white but spatially "colored," which provides the independent shocks needed to keep the autoregressive process of random fields $\{\mu(\cdot, t) : t = 1, 2, \ldots\}$ moving forward in time in a stochastic manner.

Identifiability of the decomposition (2) is maintained by writing

$$\mu(\mathbf{s}, t) = \sum_{j=1}^{p} a_j(t)\phi_j(\mathbf{s}),$$

where $\{\phi_j(\cdot) : j = 1, 2, \ldots\}$ are a complete and orthonormal sequence of deterministic spatial functions and $\{a_j(t) : t = 1, 2, \ldots\}$ is a random time series for each $j = 1, 2, \ldots, p$. The upper index p is the key to dimension reduction in the STKF and, once chosen, determines the dependence structure in the error term ν in (2). Now, because of completeness, we can write

$$\omega_{\mathbf{s}}(\mathbf{u}) = \sum_{\ell=1}^{\infty} b_\ell(\mathbf{s})\phi_\ell(\mathbf{u}),$$

and orthonormality means that equation (3) becomes

$$\boldsymbol{\phi}(\mathbf{s})'\mathbf{a}(t) = \mathbf{b}(\mathbf{s})'\mathbf{a}(t-1) + \eta(\mathbf{s}, t),$$

where $\boldsymbol{\phi}(\mathbf{s}) \equiv (\phi_1(\mathbf{s}), \ldots, \phi_p(\mathbf{s}))'$, $\mathbf{a}(t) \equiv (a_1(t), \ldots, a_p(t))'$, and $b(\mathbf{s}) \equiv (b_1(\mathbf{s}), \ldots, b_p(\mathbf{s}))'$. This equation holds in particular for locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, which allows one to isolate the vector $\mathbf{a}(t)$ on the left-hand side of a first-order autoregressive equation (Wikle and Cressie, 1997). Thus, equations

(1), (2), and (3) become

$$x(\mathbf{s},t) = y(\mathbf{s},t) + \zeta(\mathbf{s},t) \tag{4}$$

$$y(\mathbf{s},t) = \sum_{j=1}^{p} a_j(t)\phi_j(\mathbf{s}) + \nu(\mathbf{s},t) \tag{5}$$

$$\mathbf{a}(t) = H\mathbf{a}(t-1) + J\boldsymbol{\eta}(t), \tag{6}$$

where $\boldsymbol{\eta}(t) \equiv (\eta(\mathbf{s}_1,t),\ldots,\eta(\mathbf{s}_n,t))'$ and H and J are matrices depending on $\{\phi(\mathbf{s}_i)\}$ and $\{b(\mathbf{s}_i)\}$ in a manner given by Wikle and Cressie (1997).

The goal of the STKF is to predict the unseen process y based on space-time data $\{x(\mathbf{s}_i,t_i)\}$. Assuming squared-error loss, the optimal predictor of $y(\mathbf{s}_0,t_0)$ is $E(y(\mathbf{s}_0,t_0)|\{x(\mathbf{s}_i,t_i)\})$. Under Gaussian assumptions, this conditional expectation is linear in the data and can be obtained very efficiently through dimension reduction and recursive equations (Wikle and Cressie, 1997).

However, MGRA give only two STKF equations. They are

$$x(\mathbf{s},t) = \sum_{j=1}^{p} \alpha_j(t)h_j(\mathbf{s}) + \epsilon(\mathbf{s},t) \tag{7}$$

$$\boldsymbol{\alpha}(t) = P\boldsymbol{\alpha}(t-1) + K\boldsymbol{\eta}(t), \tag{8}$$

where we note that $\{h_j(\cdot)\}$ need not be othonormal and ϵ is a general space-time error field that MGRA later specialize to having dependence structure like that of ν in (5). Now, notice that if Wikle and Cressie's STKF equations (4) and (5) are combined, we obtain

$$x(\mathbf{s},t) = \sum_{j=1}^{p} a_j(t)\phi_j(\mathbf{s}) + \nu(\mathbf{s},t) + \zeta(\mathbf{s},t),$$

which yields a completely analogous equation to (7) if we write $\epsilon = \nu + \zeta$. Further, equation (8) is identical in form to equation (6). The problem with writing the reduced equations (7) and (8), as MGRA have done, is that it only allows prediction of *one* of the components of the unseen process, namely $\sum_{j=1}^{p} \alpha_j(t_0)h_j(\mathbf{s}_0)$. This explains our earlier statement that MGRA's predictions may be too smooth and also demonstrates that their mean squared prediction errors (MSPEs) will not reflect the variability inherent in predicting the scientifically meaningful process y. To put it

another way, MGRA's Kriged Kalman filter computes

$$E(\sum_{j=1}^{p} \alpha_j(t_0) h_j(\mathbf{s}_0) | \{x(s_i, t_i)\})$$

and its MSPE, whereas Wikle and Cressie's STKF computes

$$E(y(\mathbf{s}_0, t_0) | \{x(\mathbf{s}_i, t_i)\})$$

and its MSPE. The difference between the two predictors is the component $E(\nu(\mathbf{s}_0, t_0) | \{x(\mathbf{s}_i, t_i)\})$, which Wikle and Cressie show can be important.

In their Section 6.1, MGRA state that Wikle and Cressie's (1997) approach is "different" but they do not explain how. Indeed it is different, since we are predicting the quantity that is of immediate scientific interest, namely y. There is also another difference that we would like to emphasize, namely our direct modeling of the physical dynamics through equation (3). We have shown how (3) leads to (6), which is immediately comparable to MGRA's (8), but in our formulation the dynamics determine the structure of H and J in equation (6). By contrast, MGRA appear to have no *a priori* way to choose P and K in equation (8); in the example given in their Section 7, P is unconstrained and K has not been made precise (probably $K = I$ because $var(\boldsymbol{\eta}(t))$ is constrained only to be positive-definite).

We believe it is worthwhile stepping back and putting the whole notion of space-time Kalman filtering into the general context of dynamic space-time statistical modeling. The measurement equation (1) can be generalized to the *data model*,

$$x(\cdot, t) = f_d(y(\cdot, t); \boldsymbol{\theta}_d(t)), \tag{9}$$

where f_d is a stochastic functional that depends on the true process $y(\cdot, t)$ and parameters $\boldsymbol{\theta}_d(t)$. Likewise, the state equations (2) and (3) can be generalized to a *process model*,

$$y(\cdot, t) = f_p(y(\cdot, t-1), \ldots, y(\cdot, 1); \boldsymbol{\theta}_p(t)), \tag{10}$$

where f_p is a stochastic functional of past values of the true process and parameters $\boldsymbol{\theta}_p(t)$. Equation (10) models the space-time dynamics of the phenomenon, such as might be obtained by a discretization of stochastic partial differential equations derived from physical laws of temperature, pressure, and wind.

¿From a hierarchical modeling point of view, equation (9) gives the first-level probability distribution $[x|y, \boldsymbol{\theta}_d]$, equation (10) gives the second-level probability distribution $[y|\boldsymbol{\theta}_p]$, and Bayes' Theorem gives the posterior probability distribution through,

$$[y|x, \boldsymbol{\theta}] \;\propto\; [x|y, \boldsymbol{\theta}_d][y|\boldsymbol{\theta}_p], \tag{11}$$

where $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}_d', \boldsymbol{\theta}_p')'$. Then, assuming squared-error loss, the optimal predictor of $y(s_0, t_0)$ is

$$\hat{y}(\mathbf{s}_0, t_0) = E(y(\mathbf{s}_0, t_0)|x, \boldsymbol{\theta}),$$

which is the mean of the posterior distribution (11). The variance of (11) is a measure of how precise the predictor $\hat{y}(\mathbf{s}_0, t_0)$ is.

When (9) and (10) satisfy linear and Gaussian assumptions, such as for the special case (1), (2), and (3), a STKF can be obtained. Further, state-space-dimension reduction is achieved by Wikle and Cressie (1997) and MGRA by featuring dominant components of variation in (10). We shall say more about how to deal with parameters $\boldsymbol{\theta}$ at the end of this discussion.

We would like to turn now to a discussion of how other researchers' contributions can be judged with regard to MGRA's Kriged Kalman filter. The geophysical disciplines were quick to adopt the Kalman filter, certainly in the purely temporal setting but also in the space-time setting. In atmospheric sciences, the idea of using a STKF was discussed and tested in the context of numerical weather-prediction data assimilation in the early 1980s (e.g., Ghil et al., 1981) and is still an active area of research in the atmospheric and oceanic sciences (Cane et al., 1996; Miller et al., 1997; Thiebaux, 1997). Unfortunately, this literature is not well known to the statistics community. Early reference to the idea of dynamic space-time statistical models (albeit from a time-series perspective) can be found in Bennett (1979), although this important work has not been well known to the statistics community either.

If the 1980s was the decade of spatial statistics, the 1990s deservedly takes its place as the decade of space-time statistics. Several researchers independently began looking at the potential of having a dynamic temporal aspect to space-time statistical modeling. Commenting on Handcock and Wallis' (1994) Bayesian approach to geostatistical space-time modeling, Cressie (1994) suggested that a Kalman filter incorporating space and time

would be a powerful way to apply the Bayesian paradigm to modeling space-time phenomena. A little later, in an invited paper to the XVIIth International Biometric Society, Goodall and Mardia (1994) sketched the idea for an approach to space-time Kalman filtering that served as the springboard for the present paper. In the same year, Guttorp et al. (1994) considered many of the notions of the dynamic space-time model given by linear versions of (9) and (10), although they did not consider Kalman filtering. We believe that the first fully detailed, scientifically motivated, space-time Kalman filter (STKF) appeared in Huang and Cressie (1996), there applied to the problem of prediction of snow water equivalent during the winter and spring months of the western USA. The dynamic space-time structure was relatively simple, being separable in space and time. The following year, Wikle and Cressie (1997) gave the dynamic space-time model (1), (2), and (3), which implied the model (4), (5), and (6), from which was developed a fully nonstationary, nonseparable STKF. Since then, others have made contributions to the STKF but each has an inadequacy of one sort or another. We have already seen how MGRA (the current paper under discussion) combines (4) and (5) and so predicts only one of the components of the true state; consequently, the predictor they call the Kriged Kalman filter is, in general, too smooth. Berke (1998) presents a model like (7) and (8), and hence his analysis has the same shortcomings as MGRA's. Meiring et al. (1998) have a measurement-error component $\zeta(\cdot,\cdot)$ in their model that propagates in time. We know of no measurement process where this is an appropriate assumption.

Finally, we would like to discuss how the parameters $\boldsymbol{\theta}$ might be handled. In all the papers described above, $\boldsymbol{\theta}$ is either assumed known or is estimated. By "plugging in" an estimate $\hat{\boldsymbol{\theta}}$ into the posterior distribution $[y|x,\boldsymbol{\theta}]$, the predictor $E(y(\mathbf{s}_0,t_0)|x,\boldsymbol{\theta})$, and any STKF recursions, one is in effect carrying out an *empirical* Bayes analysis. Thus, all these papers, including the current one under discussion, ignore the extra variability inherent in estimating $\boldsymbol{\theta}$. A fully hierarchical Bayesian implementation puts a prior distribution $[\boldsymbol{\theta}]$ on the parameters $\boldsymbol{\theta}$ and bases all inference on the resulting posterior distribution $[y,\boldsymbol{\theta}|x]$. Wikle et al. (1998) show this to be a very powerful methodology for including complicated space-time structure, albeit with a high computational overhead. We believe that this approach represents a view into the future of dynamic space-time statistical modeling.

References

Bennett, R.J. (1979). *Spatial Time Series.* London: Pion Limited.

Berke, O. (1998). On spatiotemporal prediction for on-line monitoring data. *Communications in Statistics. Theory and Methods* **27**, to appear.

Cane, M.A., A. Kaplan, R.N. Miller, B. Tang, E.C. Hackert and A.J. Busalacchi (1996). Mapping tropical Pacific sea level: Data assimilation via a reduced state space Kalman filter. *Journal of Geophysical Research* **101**, 22599-22617.

Cressie, N. (1994). Comment on "An approach to statistical spatial-temporal modeling of meteorological fields" by M.S. Handcock and J.R. Wallis. *Journal of the American Statistical Association* **89**, 379-382.

Cressie, N.A.C. (1993). *Statistics for Spatial Data, Revised Edition.* New York: Wiley.

Ghil, M., S.E. Cohn, J. Tavantzis, K. Bube and E. Isaacson (1981). Applications of estimation theory to numerical weather prediction. *Dynamic meteorology: Data assimilation methods*, L. Bengtsson, M. Ghil, and E. Kallen, eds. New York: Springer-Verlag, 139-224.

Goodall, C. and K.V. Mardia (1994). Challenges in multivariate spatio-temporal modeling. *Proceedings of the XVIIth International Biometric Conference*, Hamilton, Ontario, Canada, 8-12 August 1994 (17pp.).

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

Handcock, M.S. and J.R. Wallis (1994). An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association* **89**, 368-390.

Huang, H.C. and N. Cressie (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics and Data Analysis* **22**, 159-175.

Meiring, W., P. Guttorp and P.D. Sampson (1998). Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Environmental and Ecological Statistics* **5**, to appear.

Miller, R.N., A.J. Busalacchi and E.C. Hackert (1997). Applications of data assimilation to analysis of the ocean on large scales. *Journal of the Meteorological Society of Japan* **75**, 445-462.

Thiebaux, H.J. (1997). The power of the duality in spatial-temporal estimation. *Journal of Climate* **10**, 567-573.

Wikle, C.K. and N. Cressie (1997). A dimension reduction approach to space-time

Kalman filtering. *Preprint Number 97-24*, Statistical Laboratory, Iowa State University, Ames, IA.

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

P. García Soidán
Universidad de Vigo, Spain
M. Febrero Bande
Universidad de Santiago de Compostela, Spain

We would like first to thank the editors for giving us the opportunity to participate in the discussion about the interesting paper by Mardia et al. This work provides a new methodology to combine in a model both the spatial and temporal components, which are derived from the data. This combination is made through two basics tools in both fields: the Kriging techniques in the spatial setting and the Kalman filter theory in the temporal one.

In what follows, we focus our comments on the application to real data sets of this general way of modelling and they are directly related to the great number of parameters that this model needs. In fact, the Kalman filter recursion conveys to the estimation of the states $\alpha(1), ..., \alpha(T)$, and the latter demands a large number of parameters to be specified. In particular, this model makes use of an initial estimate $a_{0|0}$, the true covariance $C_{0|0}$, the parameter matrix H, the transition matrix P as well as the covariance matrices Σ_ε and Σ_η.

On one hand, it is usual to assume that matrices H and P are known in the Kalman filter theory, but in practice both of them should be estimated from the data. This makes it necessary to analyze the validity of the procedure suggested under this approximation. In this sense, we propose to proceed as suggested in Young, A. & Putter, H. (1998), where it is analyzed the effect of the covariance function estimation on the accuracy of Kriging predictors. In that case, it is proved that the effect of estimation may be negligible if the joint Gaussian distributions of the process at $s_0, ..., s_n$, under both the true and the estimated covariance functions, are contiguous almost surely. Our opinion is that something similar should be valid for the Kriged Kalman filter, under appropriate conditions.

On the other hand, as pointed out in the paper, the number of parameters to be estimated can be reduced substantially by choosing matrices H and P in an adequate way; however, a large amount of data is required in order to obtain an acceptable estimation in practice. In this paper by Mardia et al, the procedure was successfully applied to 9898 data, from $T = 707$ daily observations at $n = 14$ monitored sites; nevertheless, the question remains open in case that the observations were dramatically reduced. With regard to this question, we suggest another way of reducing parameters and it could be through nonparametric estimation of covariance functions, following the ideas in Hall, P, Fisher, N.I. and Hoffmann, B. (1994) and Hall, P. and Patil, P. (1994). In both papers, nonparametric kernel estimators of the covariance function of a stationary stochastic process are described. These estimators satisfy the positive semidefiniteness property and, therefore, they can be used directly for estimation; in other words, this means that whenever it is necessary an estimate of the covariogram, this can be obtained directly data-driven, just following the latter procedure.

Finally, we would like to congratulate the authors for the interesting work developed in this paper. It is not easy to provide a methodology that deals with both space and time in the same model and combined in such an intelligent way. We find here a lot of stimulating new ideas.

References

Hall, P., N.I. Fisher and B. Hoffmann (1994). On the nonparametric estimation of covariance functions. *Annals of Statistics*, **22**, 4, 2115-2134.

Hall, P. and P. Patil (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields*, **99**, 399-424.

Young, A. and H. Putter (1998). On the effect of covariance function estimation on the accuracy of Kriging predictors. Preprint.

C. A. Glasbey

Biomathematics and Statistics Scotland, United Kingdom

This paper is a welcome addition to the burgeoning literature on spatio-

temporal models. However, no approach can hope to be fully general, because spatio-temporal data come in so many forms:

- Data may be point observations, or area (or temporal) averages, either regularly or irregularly spaced, of variables;
- Data may be uni- or multi-variate, and may be far denser in space than time, or conversely;
- Variables may be continuous, binary or categorical (ordered or unordered), in discrete or continuous space and/or time;
- Variables may, or may not, be stationary in space and/or time.

In this paper, the model is developed in the first instance for data that are univariate point observations, positioned irregularly in space but regularly in time, and which are far denser in time than in space. The variable (after a transformation) is assumed to be a Gaussian random field in continuous space and time, and to be stationary in both space and time. Some of these aspects of the model are less restrictive than others. For example, the same methodology could be used for multivariate observations which are positioned regularly in space and which are denser in space than time. Also, in their discussion the authors mention extensions to non-Gaussian, block-averaged and seasonally-dependent variables. However, alternative spatio-temporal models will be needed for data which are irregularly sampled in time, for image sequences, categorical data and for general non-stationary processes.

The proposed model can be interpreted as a singular value decomposition (SVD) of the data matrix, x, but with coefficients h and α that have been smoothed in space and time, respectively. In comparison, Glasbey (1992) simply used a SVD to interpret spatio-temporal variation in solar radiation. The advantages of smoothing are that it reduces the variances of the SVD coefficients and makes it possible to interpolate between observations.

In my opinion, two particularly welcome features of the model are:

1. It reduces to known classes of time series and spatial models, if only a single site or a single time is considered;

2. The space-time correlations are non-separable, i.e., the correlation at spatial dispacement s, temporal dispacement t, is not simply the product of the spatial correlation at dispacement s and the temporal correlation at dispacement t.

The first issue is important, because most spatial data sets are samples at single instants in time of spatio-temporal processes in the real world. Similarly, many time series are samples at single locations of spatio-temporal processes. Therefore, it is desirable for a spatio-temporal model to generalise the existing spatial and time series models for these samples. As regards the second issue, in the proposed model, auto- and cross-correlations are sums of p exponentials, plus a white-noise contribution from ϵ, and are therefore not space-time separable. In contrast, separability is assumed in many models, for mathematical convenience, usually without having any basis in the data. The assumption may appear weak, but can have major consequences, one of which is discussed by Glasbey (1995).

Much remains to be done in spatio-temporal modelling. Spatio-temporal data sets are typically very large: more experience is needed in assessing how complex a model is needed to summarise them, and more tools are needed for model identification and validation.

References

Glasbey, C.A. (1992). A reduced-rank regression model for local variation in solar radiation. *Applied Statistics*, **41**, 381-387.

Glasbey, C.A. (1995). Imputation of missing values in spatio-temporal solar radiation data. *Environmetrics*, **6**, 363-371.

John T. Kent
Univeristy of Leeds, United Kingdom

The paper proposes a sophisticated statistical model to describe spatial-temporal data. The key idea is to model the bulk of the spatial pattern in terms of a small number of basis fields $h_j(s),\ s \in \mathbb{R}^k$, whose coefficients $\alpha_j(t)$ vary through time according to a cross-correlated set of stochastic

processes. Several aspects of the model, both mathematical and practical, merit further discussion.

1 Kriging vs. Kalman filter

The basic model of the paper, the strong ST–GSS model of Section 2, takes the form

$$\mathbf{x}(t) = H\boldsymbol{\alpha}(t) + \boldsymbol{\varepsilon}(t), \quad \boldsymbol{\alpha}(t) = P\boldsymbol{\alpha}(t-1) + \boldsymbol{\eta}(t), \tag{1}$$

where the $\{\boldsymbol{\varepsilon}(t)\}$ and $\{\boldsymbol{\eta}(t)\}$ are independent sequences of iid random vectors. The introduction of the paper gives Kriging and the Kalman filter equal billing, but in an important sense the autoregressive model underlying the Kalman filter is much more fundamental. Suppose the parameters $H(n \times p)$, $P(p \times p)$, $\Sigma_\varepsilon(n \times n$ positive definite) and $\Sigma_\eta(p \times p$ positive semi-definite) are known. Provided $n \geq p$ and H has full rank, (1) can be linearly transformed to remove the spatial interdependence between the n data sites, and to simplify the relationship between $\mathbf{x}(t)$ and $\boldsymbol{\alpha}(t)$, yielding

$$\mathbf{x}_1'(t) = \boldsymbol{\alpha}'(t) + \boldsymbol{\varepsilon}'(t), \quad \mathbf{x}_2'(t) = \boldsymbol{\varepsilon}_2'(t), \tag{2}$$

$$\boldsymbol{\alpha}'(t) = P'\boldsymbol{\alpha}'(t-1) + \boldsymbol{\eta}'(t). \tag{3}$$

Here $\mathbf{x}'(t) = (\mathbf{x}_1'(t)^T, \mathbf{x}_2'(t)^T)^T$ and $\boldsymbol{\varepsilon}'(t) = (\boldsymbol{\varepsilon}_1'(t)^T, \boldsymbol{\varepsilon}_2'(t)^T)^T$ (each partitioned into pieces of sizes p and $n-p$, respectively), and $\boldsymbol{\eta}'(t)$ are the linear transforms of $\mathbf{x}(t), \boldsymbol{\varepsilon}(t)$, and $\boldsymbol{\eta}(t)$, respectively, and $\boldsymbol{\varepsilon}'(t)$ and $\boldsymbol{\eta}'(t)$ now have uncorrelated components. Thus spatial autocorrelation has been removed from (2). In (3) the fact that P' is generally not diagonal leads to interdependence between the components of $\boldsymbol{\alpha}'(t)$. Further $\mathbf{x}_2'(t)$ does not depend on $\boldsymbol{\alpha}'(t)$ and so is available for diagnostic purposes. Of course the spatial aspects of the model become more prominent for prediction and in the definition of H.

2 Principal fields

The systematic components of the model $\{h_j(s),\ j = 1, ..., p\}$, consist of trend fields and principal fields. The purpose of the principal fields is to get a set of fields which (a) vary from low to high frequency variation across the part of the region containing the data sites, and (b) die away as s moves

far from the data sites. If the sites were equally–spaced in a rectangle in $\mathbb{R}^k$ with periodic boundary conditions, then the principal fields would reduce to the usual trigonometric fields from Fourier series, and hence have a simple explicit form. (These fields satisfy (a), but (b) is not relevant.) Thus in the nonperiodic case it is worth considering whether it is possible to construct simple set of explicitly-defined fields which fulfill the requirements of (a) and (b). One possibility is to use wavelets, which might lead to more interpretable state vectors $\boldsymbol{\alpha}(t)$.

3 Interpretation

Suppose a plume of pollution crosses the study region. The principal fields of the paper possess no explicit mechanism to model or recognize such behaviour. In the periodic case with trigonometric fields, such behaviour would manifest itself by a constant amplitude over time at each frequency and by phase angles that vary over time in a common manner across frequencies. Perhaps something of this interpretability can be preserved using wavelets. Alternatively, if enough spatial data are available, it might be fruitful to consider a more physically-based model incorporating effects such as wind velocity, diffusion, and sources of pollution.

4 Model validity

What is the ultimate purpose behind the statistical models used in this paper — is is merely to provide pretty maps, or is it intended to provide some useful statistical confidence statements about predictions at new sites. Predictions at new sites involve faith about how the key aspects of the model (e.g. the form of spatial covariance function and the choice of common fields) can be extrapolated from the data sites. Cross-validation seems to be an essential tool in justifying this extrapolation. But there is a also a need for extra care in case the part of the study region without data sites differs in some fundamental way from the part of the study region containing the data sites (e.g. mountainous vs. flat, or urban vs. rural).

Ana F. Militino
Universidad Pública de Navarra, Spain

It is a great pleasure to have the opportunity to comment this interesting paper. The authors have provided an stimulating and readable account of an approach to the problem of spatial-temporal modeling via Kriging Kalman filter.

Firstly, they introduce the basic concepts and the formal structure of the spatio-temporal model, called the Kriged Kalman Filter (KKF) model, following with a detailed explanation of the applied methodology used in the estimation procedure. Finally, they illustrate the results with a pollution problem.

The goal is not easy at all. Its great complexity has made that only outstanding statisticians contribute to this field. Spatio-temporal modeling has been developed in the statistical literature for different applications and with different procedures in the estimation process. The applications have been focused on the environmental field, that is why there is a huge variety of problems and a diversity of associated solutions. Here, in this paper I find a sensible way of tackling with this kind of model. It is a perfect combination of two classical techniques already used in a separate way in the estimation procedure. Kriging in the spatial context and Kalman filter in the temporal one. Both, coming from the engineering field, but extensively used in other areas.

Exploring this paper, I have been thinking about several questions which I would like to ask to the authors.

Kriging and Kalman filtering assume that the covariance structure is known. Usually, this is not the case and a subestimation of the prediction error is made. Moreover, as far as the spatio-temporal modeling in environmental studies is concerned, most spatial covariance structures are not stationary over the spatial scales of interest. Sampson and Guttorp (1992) indicate that in studies of dispersion of atmospheric pollutants, landscape or topography affect weather patterns and, thereby, the transportation of pollutants. Handling with observations depending on both time and spatial locations, how one can check these desired properties or how it is possible to overcome them if they fail? Even more, the combination of spatio-temporal stationarity seems a very difficult question. By the other hand, neither the Kriging nor the Kalman filter need assumptions of normality

to be applied, but their optimality virtues are extremely linked with this assumption. Then, it is natural to think at least in a necessary assumption of normality to guarantee the joint optimality.

With regard to the estimation process, the number of parameters involved in the spatial-temporal model is so large that we could find serious problems of identification and estimation, above all when the estimation procedure is a regressive one where the errors are accumulative.

The maximization of the likelihood function is based on the EM algorithm, which is a standard tool in the statistical repertoire. Here, I find the well-known quibbles: the EM algorithm does not guarantee convergence to the global maximum when there are multiple maxima and in this case, the estimate obtained depends on the initial value. It also has a slow rate of convergence and the most awkward matter, the algorithm does not produce estimates of the covariance matrix of the MLEs. Perhaps it is for this reason that we can not find standard errors in this paper, although there are additional methods that can be integrated into the EM computational scheme.

I would like to finish my comments saying that overall I enjoyed reading the paper. It made me feel that there is still a possibility of improving the knowledge about spatio-temporal modeling using traditional tools.

References

Sampson, P. D. and P. Guttorp (1992). Nonparametric Estimation of Nonstationary Spatial Covariance Structure. *Journal of the American Statistical Association*, **87**, 108-119.

Michael L. Stein
University of Chicago, U.S.A.

This research was supported in part by the National Science Foundation grant DMS 95-04470.

1 Space-time modelling

Developing appropriate stochastic models and statistical methods for space-time processes is one of the great open challenges in statistics and this paper (MGRA from now on) is a welcome addition to this rapidly expanding literature. In my discussion I would like to explore some conceptual problems that arise in trying to merge the Kalman filter with Kriging.

The Kalman filter provides a computationally efficient method for updating past predictions of the state of a system. These predictions are optimal when the process possesses an appropriate Markov property. For example, in the simplest case, one assumes that the observation error process $\epsilon(t)$ is independent in time, the innovation process $\eta(t)$ is also independent in time and the two processes are independent of each other. In this setting, both $\alpha(t)$ and $(\mathbf{x}(t), \alpha(t))$ are vector-valued autoregressions of order 1 and hence first-order Markov chains. Efficient algorithms for computing optimal predictions also exist when $\epsilon(t)$ and $\eta(t)$ are assumed to be independent vector-valued autoregressions of order 1 (Chui and Chen, 1991, Chapter 5).

Equations (2.2)–(2.5) give a clever model possessing a Markov property in time that makes it possible to calculate optimal linear predictions at the present or a future time using only data from the present time and a prediction of the state vector at the present time, which can be computed efficiently using a Kalman filter. This computational efficiency comes at the cost of having a model with space-time covariances of a highly specific form. Indeed, I will argue that MGRA's model is only modestly related to what would generally be called Kriging.

In (3.1)–(3.5), MGRA give the usual Kriging predictor for a spatial process $x(\mathbf{s})$ with one important difference: in Kriging as it is generally used, $\mathbf{x}$ is the entire vector of n observations $x(\mathbf{s}_1), \ldots, x(\mathbf{s}_n)$ and there is no selection of a lower-dimensional subspace of the bending energy matrix B as a basis for prediction. As MGRA point out in Section 8, using only selected principal fields does entail a loss of spatial information. However, there is another important way in which their approach differs from Kriging. Rewriting (3.3) slightly, the Kriging predictor $\hat{x}(\mathbf{s})$ can be viewed as a function of $\mathbf{s}$:

$$\hat{x}(\mathbf{s}) = \sum_{i=1}^{n} w_i \sigma(\mathbf{s}, \mathbf{s}_i) + \mathbf{f}(\mathbf{s})^T \beta,$$

where the w_is and β are determined by the trend matrix F and the covariance matrix Σ_ζ of the observations and are independent of $\mathbf{s}$. MGRA exploit this formulation of the Kriging predictor to motivate their approach. Specifically, similar to Kriging, they use the functions $\mathbf{f}$ and σ to define their p common fields. On the other hand, unlike the Kriging approach, their w_is and β are not determined by F and Σ_ζ but are taken to be a further set of parameters α to be chosen using a Kalman filter. We can now see why MGRA could not take their normative sites to be all of the observations and use all of the resulting principal fields ($n = m = p$). They would then always be able to choose $\alpha(t)$ to perfectly fit the observations at time t and would end up with fits that do not smooth in space and do not use observations from past times in the prediction. MGRA avoid this problem by choosing the dimension of $\alpha(t)$ to be much smaller than n, but the resulting procedure strikes me as having more in common with other methods that make use of common fields (such as empirical orthogonal functions or principal oscillation patterns (von Storch et al., 1994)) than it does with Kriging.

I also have some general concerns about the approach for choosing the principal fields. The idea of selecting the eigenvectors associated with the small eigenvectors of the bending energy matrix B is a sensible way of capturing the large-scale spatial variations in the process. A necessary implication of this approach is that small-scale spatial variations, which can now only be captured in the $\epsilon(\mathbf{s}, t)$ term, cannot have any time dependence. While it may often be plausible that small-scale spatial variations show little or no time dependence, it is at least logically possible to have a pollutant level with large differences over small distances that persist over several days. It would be worthwhile to have diagnostics that could detect this sort of departure from MGRA's model.

I would like to suggest another approach to modeling processes in continuous space and discrete time that I would argue is more in the spirit of Kriging. For simplicity, suppose that trends in both space and time have been removed from the data so that we can assume $Ex(\mathbf{s}, t) = 0$ for all $(\mathbf{s}, t)$. As in (2.2), write $x(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \epsilon(\mathbf{s}, t)$ and assume $\epsilon(\mathbf{s}, t)$ satisfies (2.5), i.e., has no correlations across time. Instead of using (2.3) as the model for $\mu(\mathbf{s}, t)$, consider modeling the function $\mu(\cdot, t)$ on S by $\mu(\cdot, t) = P\mu(\cdot, t-1) + \eta(\cdot, t)$, where P is a linear operator mapping continuous real-valued functions on S onto this same space of functions. Furthermore assume $\eta(\cdot, t)$ has no dependence in time. Since the state of the

system at time t is arguably the function $\mu(\cdot, t)$ on the domain S, one can make a case that this model is the natural generalization of the Kalman filter to this setting. Of course, the state of the system is now infinite-dimensional whenever, as is generally the case, S is not finite, but this does not change the fundamental nature of the model. Suppose, for simplicity, that P is known. A plausible way to proceed would then be to estimate the covariance structures for $\epsilon(\mathbf{s}, t)$ and $\eta(\mathbf{s}, t)$, perhaps by assuming they are both isotropic processes in space. While I find this model to be more natural than the one proposed by MGRA, it does have an important drawback. Specifically, the implied model for the observations of this space-time process at a finite number of sites will generally not have a structure that permits the direct application of a Kalman filter to calculate optimal linear predictions unless P has some special form, e.g., P is diagonal. However, since the process $\mu(\cdot, t)$ is Markov in time, the resulting model for the data can be viewed as an example of a hidden Markov model and I would imagine that someone sufficiently clever could find a way to compute likelihoods and predictions accurately and efficiently under it.

2 The example

I have some general and specific comments about the example treated in MGRA. My first comment concerns the use of empirical covariance functions as a way of selecting models and estimating parameters for the covariance functions of spatial data. In my experience, such plots can be highly misleading. The main problem is that many of the points in such a plot are often highly correlated so that fitting such a plot by eye can yield poor parameter estimates. The fact that the observation sites are highly clustered only makes this problem worse. While MGRA do end up estimating b using maximum likelihood, I would encourage them to estimate all of the parameters of the covariance function via the likelihood and, furthermore, to look at the likelihood surface as a way of learning about the uncertainty in these estimates.

A second concern I have with MGRA's methodology is their choosing the normative sites to be the same as the actual sites. Given that most of the sites are in one of two clusters, this choice practically guarantees that the first principal field will essentially be a difference between these two clusters. I would find it more natural for the principal fields to represent

the pattern of variations throughout the region of interest rather than those at the observation sites. Thus, I would suggest taking the normative sites to be evenly spaced throughout the region of interest, whether or not the actual observations are evenly spaced. It would be interesting to see how much MGRA's predictions would change if the normative sites were evenly spaced. I would guess that predictions at locations that are not very near any of the observations would show a bit more variability than was obtained in Figure 9.

A modification of MGRA's model worth considering is to include a random process that varies in space but not in time. Figure 2 suggests that the northwestern corner of the region under study tends to have the lowest sulfur dioxide levels throughout the study period, suggesting the need for a term that varies only in space. While leaving out such a term may have only a minor effect on predictions of daily sulfur dioxide levels, Stein and Fang (1996) demonstrate that this omission can lead to dramatically underestimating the uncertainty of predictions of long-term averages.

The residual plots given in Figure 8 are a useful diagnostic tool. For example, these plots are more effective at identifying outliers than the plots of the raw data in Figure 2. In particular, sites 7 and 12 each show one strongly negative residual. Are these outliers severe enough that they have a large impact on the overall analysis? If so, since the negative residuals correspond to unexpectedly low pollution levels, which are generally of much less interest than unexpectedly high levels, it would be worth considering employing robust methods that would reduce the influence of these observations.

Finally, a rather different direction for modeling this and similar datasets would be to combine statistical and physical models. Since sulfur dioxide is a primary pollutant, physical models that make use of information about emissions sources and wind patterns in the region ought to be able to produce decent representations of the spatial patterns in sulfur dioxide on a daily basis. One could then combine this analysis with some stochastic modeling by assuming, for example, that the difference between the logarithms of the actual sulfur dioxide levels and those obtained from the physical model are a space-time random process. I would expect such a model would produce better predictions than models using either just a physical approach or just a statistical approach. Furthermore, the differences between the observations and a physical model prediction may have a

simpler structure than the observations themselves, for example, by being more nearly stationary and isotropic in space. To carry out such a combined modeling approach would undoubtedly take the active collaboration of statisticians and atmospheric scientists and I can only see good things coming out of such an endeavor.

References

Chui, C. K. and G. Chen (1991). *Kalman Filtering with Real-Time Applications.* Berlin: Springer-Verlag, second edition.

Stein, M. L. and D. Fang (1996). Contribution to the discussion of "Ozone exposure and population density in Harris County, Texas." *Journal of the American Statistical Association,* **92**, 408–411.

von Storch, H., G. Bürger, R. Schnur and J. von Storch (1995). Principal oscillation patterns: a review. *Journal of Climate,* **8**, 377–400.

Rejoinder by K.V. Mardia, C. Goodall, E. Redfern and F.J. Alonso

We are grateful to each of the discussants for taking time to comment, in spite of the fact that comments were requested during the summer period. There are a wealth of ideas in the written discussions. We were also fortunate to present the paper to the 24th National Conference of the Spanish Statistical and Operational Research Society in Almería and the oral discussion was also illuminating. Kanti Mardia and Ed Redfern would like to take this opportunity to thank the organizers for their hospitality; the setting of the conference could not have been more idyllic.

The comments can be grouped into the following themes: (i) General Modelling and KKF, (ii) Kriging vs Kalman Filter, (iii) Model Selection: Covariances, and Covariance Estimation, (iv) Model Selection: Common Fields, (v) Implementation Issues, (vi) Diagnostics, (vii) The Illustrative Example, (viii) History and Future of the Subject.

1 General modelling and KKF

All of the discussants have pointed out various approaches to modeling spatial-temporal data. Most of the models, including our KKF model, resemble the singular value decomposition in the sense of separating out temporal-invariant spatial patterns and spatial-invariant temporal patterns. For the KKF these are respectively a basis for Kriged surfaces - the common fields - and the multivariate general state space model. The main points of our KKF model are: (a) The observations are linear combinations of spatial fields, called common fields, at each time, (b) The coefficients of the linear combinations comprise the state vector in a general state model, (c) The common fields include trend fields and principal fields for the so-called normative sites, and together comprise a basis for Kriging solutions for data at the normative sites, (d) Estimation is via maximum likelihood, (e) The problem of simultaneous estimation in space and time is reduced to recursive estimation, using the Kalman filter and MLE, of a multivariate general state space model, (f) The approach to spatial-temporal data is one of exploratory modeling, using a broad and flexible class of models, (g) We offer here one complete strategy among many; there is much scope for further development, including specialization. Among various strategy described in Goodall and Mardia (1994), we have adopted here one single but full strategy.

Modeling spatial temporal data is necessarily a complex undertaking. As Dr. Glasbey points out, at a fixed spatial location we have a univariate time series, at multiple fixed spatial locations a multivariate time series, and at a fixed time a spatial modeling problem. Beyond that are the problems of temporal and spatial covariances appearing in the same model, in a non-separable way (as underscored by Dr. Glasbey and Professor Cressie and Dr. Wikle). Fortunately help is at hand: Kriging and Kalman filtering each have associated with them a wealth of practical experience and extensions, which can be incorporated into KKF. We are grateful to several discussants, Professor Militino, and Professors Soidan and Bande, for underscoring this, and to many discussants for sharing their experience and suggestions.

The decomposition given by Cressie (1993, equation 3.1.2), of a random field into several components of variation (at different spatial scales) is fundamental to our understanding of what contributes to the error term $\Sigma_{\boldsymbol{\varepsilon}}$ in the observation equation, modified in the spatial temporal setting

by the (strong ST-GSS) assumption that all temporal dependence is in the state vector. Professor Cressie and Dr. Wikle's discussion (hereafter CW) rightly reminds us that the appropriate components of spatial variation, their ν, can be included in spatial prediction, as we state at the very end of Section 5 and give explicitly in (6.2) and (6.3), and in variance calculations, as we give in (6.4). Thus our opinion is that in general the KKF does not give over-smoothed prediction. Indeed, in our illustrative example we could have used more complicated covariance assumptions to highlight this point but we decided to select a particularly simple covariance assumptions for a clear exposition.

Spatial structure that is independent from one time point to the next is qualitatively different from spatial-temporal structure. If the time interval between observations is diminished, then we expect to find temporal correlations appearing between spatial structures (the decomposition again, but now in time). Whereas purely spatial prediction should include the spatial structure, predictions forward in time include only the spatial-temporal part. The spatial-temporal fit at every time point is an important and interesting quantity.

One way to describe the CW model is to say that they include errors-in-variables in the spatial component. Their definition of smoothness differs, for example, in terms of the second order derivatives in the spline.

2 Kriging vs Kalman Filter

Professor Kent and Professor Stein suggest that the Kriging component in the KKF model is not as fundamental as the Kalman filter. We are grateful to them for this observation but would give the two methods complementary and thus more equal roles. First, the role of Kriging is fundamental to spatial prediction and formulation of H. Second, the following thought exercise may help underscore their respective roles in KKF. Suppose that the state innovations, $\boldsymbol{\eta}$, have very large variance. Then the KKF model becomes separate weighted generalized least squares predictions at each time, that is, separate Kriging predictions provided that the common fields are a basis for Kriging predictors. We can write down a set of Kriging models with covariance Σ_ζ taken equal to $\Sigma_{\boldsymbol{\varepsilon}}$. We may look for this analytically by comparing the Kriging equations in Section 3 to the Kalman prediction and updating equations in Section 5.1 when the variance of $\boldsymbol{\eta}$ is very large.

The state equations impose constraints on the state $\mathbf{a}_{t|t}$, that is, the "Kriging solution", in proportion to the relative magnitudes of $\Sigma_{\boldsymbol{\varepsilon}}$ and $\Sigma_{\boldsymbol{\eta}}$. We are reminded of Meinhold and Singpurwalla's (1983) Bayesian view of the Kalman filter. The converse is also true: the common fields derived from the spatial model constrain the "Kalman filter" parameter matrix H. These constraints are weakest when the maximum-likelihood estimates of H in a multivariate general state space model coincide with the H obtained using common fields, possibly to within a linear transformation. However, neither extreme situation is desirable, as they involve loss of either the Kriging or of the Kalman filtering part of KKF.

Professor Stein's alternative, to model the mean function resembles Green and Titterington (1988), who introduce a spatial-temporal Kalman filter in which the state vector is taken to be the spatial field.

3 Model selection: covariances, and covariance estimation

Our Table 1 shows various stages of our model selection. One of the key problems is the selection of covariance functions for principal fields and for the observation equations. Professor Soidan and Professor Bande suggest using non-parametric estimation of these covariance functions. Some details regarding combining non-parametric methods with parametric methods should be worked out, for example, the cross-covariance functions. We look forward to reading the unpublished paper of Young and Putter (1998) quoted by Professors Soidan and Bande. While the predictor may be robust to estimation (except when there is significant extrapolation), standard errors are critically dependent on the estimated covariances.

4 Model selection: common fields

The key to combining Kriging and Kalman filter into the KKF model is the use of common fields. We have selected a particular implementation, the use of trend and principal fields, but other choices are appropriate, as various discussants suggest. Typically there is a dimension reduction ($p < m \leq n$), but, to address Professor Stein's concern, we are motivated by modeling parsimony rather than by statistical necessity: in GSS modeling, including, e.g., a univariate ARMA model, we regard the state vector to be a

hidden process which does not contribute directly to the total parameter count. The simultaneous minimization of innovation and measurement errors ensures that even without dimension reduction the fits are not exact. In practice the choice of common fields does resemble empirical orthogonal functions: do the T separate Kriging solutions at each time point lie in a subspace?

The choice of principal fields with large scale or with small scale variation (or some combination of the two), are advanced as possible convenient data reduction techniques that might or might not be sensible in any situation. In the example, a model with the large scale variation principal fields gave a far greater reduction in negative log likelihood than did a model with the small scale fields. Another technique is to choose a small set of the normative sites, and all principal fields. The results with data were not quite as good as when there are more sites; in fact there is some indication to use a small number of fields defined on a grid of normative sites.

The discussants' comments on both principal fields and common fields generally are immensely helpful in advancing the set of alternatives, including regular spaced normative sites (Professors Stein and Kent), concentration of normative sites where small scale variation is most apparent, Fourier surfaces and trigonometric fields (Professor Kent), and wavelets (Professor Angulo and Professor Kent). The flexibility of wavelets in modeling patterns at different resolutions is clearly valuable. Trigonometric fields are advantageous in some situations, and Professor Angulo rightly proposes specifying the contrasts in advance. These alternative choices of common fields do not create any new problems, but their behaviour in practice will need a detailed investigation.

The smooth time evolution of the common fields through a lag window, and the deeper notion of letting the bending energy matrix change dynamically, are excellent ideas from Professor Angulo. Goodall and Mardia (1994) also remark on letting different parts of the model vary at different temporal scales. The same statement can be made spatially, for example, the choice of common fields and general state space model can vary slowly in space.

Returning to discussion of principal fields, the selection of principal fields with large scale variation has parallels to principal components regression, the drawback to which is, of course, that variation in the response need not follow principal axes of variation in the predictors. For a fixed

set of sites, linear combinations of the principal fields (computed for these sites) provides optimal, Kriging, prediction separately at each time point. The exploratory selection of normative sites to define the common fields becomes reasonable when we consider the variable sets of sites and data types (images) and the probabilistic constraints on the linear combinations imposed by the system equation. Other approaches include informing the choice of common fields with physical modeling (Professor Cressie and Dr Wikle), and using exploratory modeling at one step remove, by exploratory modeling of the residuals from fitting a physical model (Professor Stein). CW and Stein differ in that CW include statistical and physical aspects in a single model.

There are limitations to the use of common fields, including Kent's moving plume example, which we allude to, and Goodall and Mardia (1994) discuss explicitly, proposing use of a nonlinear Kalman filter to deal with that case, which is then taken up in Cressie and Wikle's discussion.

5 Implementation issues

We have used the EM algorithm which as pointed out by Professor Militino suffers from convergence problems. Also, it does not provide standard errors of the estimates. Further, since we are using a recursive method, the errors are cumulative in time. We are presently exploring full maximum likelihood estimation of the covariance and Kalman filter parameters using a Monte-Carlo Markov Chain, which has the advantage of providing measures of uncertainties. Also, the procedures adopted by Diggle *et al.* (1998), and the hierarchical approach of Cressie *et al.*, have good potential. A Bayesian framework, not however utilized in the paper, can be used to incorporate prior information and has implementation advantages.

6 Diagnostics

Professor Militino rightly points out that environmental data suffers from non–stationarity in general. Diagnostics to detect departure from the KKF model will be worthwhile, especially to detect large differences over small distances, persisting over several days as commented by Professor Stein. Goodall and Mardia (1994) have made some suggestions.

7 The illustrative example

Various specific points on the example have been commented on by Professor Stein which have not been answered above. The comment about using the empirical covariance functions agrees with our experience. In this case we estimated only the parameter b to save on computation. Other values of the parameter p were tried but did not result in an improved fit. The fact that the initial value for b of 0.1 was so far from the final value of 1.3 confirms the inadequacy of the estimated covariogram (averaged over all time points) as a diagnostic tool. The issue of stationarity over time is also a problem here. Three strategies suggest themselves for normative sites: using (1) a subset of observed sites, (2) the full set of observed sites, and (3) a grid of points covering the region in which data is observed. Strategy (2) as used in the example produces the best fit and minimizes the assumption about the behaviour away from these sites as the surface damps to the underlying trend surface. Using a grid of points would introduce more variability away from the observed sites as is suggested. Omitting the two observations in sites 7 and 12, mentioned by Professor Stein, does not have a dramatic effect on the model. However the point about using robust methods to reduce the influence of discordant observations is taken and merits further research.

8 History and future of the subject

It is most helpful of Professor Cressie and Dr. Wikle to give us a historical perspective of the subject. Goodall and Mardia (1994) made such an attempt but the description by CW is much broader. There will be always some omissions for such a fast developing area. For example, how does Dr. Glasbeys work fit into this scheme? Mardia and Goodall (1993) have another perspective. Also what about the area of spatial-temporal modelling for point patterns. We agree with Professor Cressie and Dr. Wikle that hierarchical modelling has a prosperous future. We wholeheartedly believe in common with all the discussants that we will see great strides in this subject of spatial-temporal modelling and that this paper will be one of many in stimulating interest in the area. We expect to see new models, accompanying software, diagnostics, and visualization techniques. Spatial point patterns that evolve in time is another large area. We have focussed

only on continuous space and discrete time but there is scope for various extensions as described in Goodall and Mardia (1994) and the last Section of this paper. What we have done is just scratch the surface!

Turning to KKF, we believe that specific, detailed strategies may evolve to deal with different real-world situations and classes of spatial temporal data. Goodall and Mardia (1994) enumerate 9 ingredients of spatial temporal modeling, which we believe that KKF approaches. These include spatial, temporal, and multivariate parts to the model. Two more are image data, and the possibility that data occur at irregular intervals of time. Dr. Glasbey rightly points out that KKF is oriented towards data that are spatially irregular and temporally regular, though data that are occur irregularly in time can be accommodated using a hopefully-coarse discretization of time and 1-, 2-, or more-step predictions, incorporating data however (images) and whenever they occur. A further point is that the GSS model allows short series also.

We agree entirely with the comments of Professor Stein and Professor Kent, also a theme through Professor Cressie and Dr. Wikle's discusion, that one should incorporate physical aspects of the model when there is a priori information. A strong link with practitioners should improve the modelling process. KKF is a flexible, tractable model that can be used for exploratory modeling. Thus we would not want to restrict our choices of P and K a priori, which CW appear to do with their equation (6), depending on the ϕ and b's. We are reassured that a scientific explanation can be attached to our models, but wish to remain flexible.

At casual reading CW's historical remarks may be misleading, so a few comments are in order here. We can identify (at least) three strands in spatial temporal modeling, namely (1) geophysical research and its statistical correlates, for example Guttorp and Sampson (1994), (2) spatial-temporal Kalman filter and the Kriged Kalman Filter, and (3) Bayesian and hierarchical Bayesian space time modeling. The spatial temporal Kalman filter described by Goodall and Mardia (1994) contains all of the essential ideas of KKF. We have used the new name KKF for our STKF in order to better highlight this model's origins in Kriging and the Kalman filter. The discussion has led us to emphasize in this rejoinder just how complementary the two techniques are in KKF. Goodall and Mardia include much detail not found in this paper, but do not include a full worked out example.

Goodall and Mardia (1994) were motivated by earlier work on spatial

and multivariate spatial-temporal modeling (Mardia and Goodall (1993)). Pushed by the forthcoming International Biometrics Congress, we saw, in early 1994 while Goodall was visiting Bristol, from Janacek and Swift's (1993) clear exposition of the general state space model (see review in *JASA*, Goodall (1994)), how to build the STKF. The use of principal fields came from Bookstein (1989), Kent and Mardia (1994), and Mardia *et al.* (1996). This was now a spatial temporal Kalman filter (STKF) in concrete form for the first time, with novel aspects in spatial temporal modeling that re-appeared in Wikle *et al.* (1997) and Wikle and Cressie (1997). Green and Titterington (1988) turned out to be a related approach, but from the point of view of modeling the spatial field as the state vector.

To sum up, KKF is a versatile technique for exploring and modeling spatial-temporal data. The practitioner need have expertise in Kriging and in multivariate time series modeling via the Kalman filter. To do both is a significant undertaking, which goes to partially explain the time lag between the first publication and this publication "with example". We are grateful to all the discussants for their interesting input, and for sharing their accummulated expertise, some of it directly with STKF, in this forum.

References

Goodall C.R. (1994). Review for *Journal of the American Statistical Association.* of Janacek, G. and Swift, L. (1993). *Time Series: forecasting, simulation, applications.* Ellis Horwood, New York.

View publication stats