Model comparison and selection for stationary space–time models

H.-C. Huang\textsuperscript{a}, F. Martínez\textsuperscript{b,1}, J. Mateu\textsuperscript{c,*1}, F. Montes\textsuperscript{b,1}

\textsuperscript{a}Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan
\textsuperscript{b}Department of Statistics and O.R., Universitat de València, E-46100 Burjassot, Spain
\textsuperscript{c}Department of Mathematics, Universitat Jaume I, Campus Riu Sec, E-12071 Castellón, Spain

Received 1 November 2004; received in revised form 1 December 2005; accepted 27 July 2006
Available online 22 August 2006

Abstract

An intensive simulation study to compare the spatio–temporal prediction performances among various space–time models is presented. The models having separable spatio–temporal covariance functions and nonseparable ones, under various scenarios, are also considered. The computational performance among the various selected models are compared. The issue of how to select an appropriate space–time model by accounting for the tradeoff between goodness-of-fit and model complexity is addressed. Performances of the two commonly used model-selection criteria, Akaike information criterion and Bayesian information criterion are examined. Furthermore, a practical application based on the statistical analysis of surface shortwave radiation budget (SSRB) data is presented.

Keywords: AIC; BIC; Kalman filter; Maximum likelihood; Nonseparable and separable spatio–temporal covariance function; Surface shortwave radiation budget (SSRB)

1. Introduction

Many environmental processes, such as groundwater contaminants, air pollution patterns, air–water and air–soil energy exchanges, are characterised by variability and uncertainty that may result from ontologic factors and epistemic limitations. The former are due to the inherent complexity of the natural systems, whereas the latter are associated with incomplete information. Classical statistical approaches often fail to provide a sufficient description of a process evolution across space and time. Spatio–temporal random fields are used to represent correlations between fluctuations in the composite space–time domain. Modelling the effects of fluctuations with suitable covariance functions can improve our ability to characterise and predict space–time variations in various natural systems. Thus, the introduction of models able to integrate spatial and temporal behaviours at different scales constitutes an important topic in probability and statistics in the last decade.

The importance of simultaneously studying spatial and temporal aspects of processes is well-known. Stochastic models describing how processes vary across space and time are essential to the application of statistics to a wide...
range of environmental applications. However, the structural analysis of such processes is more difficult than for spatial or temporal processes. Estimating and modelling the correlation of a space–time process are principal goals in geostatistical analysis. In the last years, several efforts have been made for the construction of valid spatio–temporal covariance structures, making the spatio–temporal modelling an important area that is one of rapid growth at the moment, with various applications in environmental science, geophysical science, biology, epidemiology and others. The reader is referred to the comprehensive reviews of Kyriakidis and Journel (1999) and Kolovos et al. (2004).

Dimitrakopoulos and Lou (1994) proposed some geometrically anisotropic models for spatio–temporal models. Guttorp et al. (1994) proposed some separable covariance structures, obtained with the product of a spatial and a temporal covariance. In a similar context of separability, Rohuani and Hall (1989) proposed the so-called sum-variogram. Other approaches assume that the drift could entirely catch the temporal variability, and consider the spatio–temporal process to be temporally independent (Sampson and Guttorp, 1992). In the nonseparable context, several efforts have been made in order to obtain a class of spatio–temporal covariances which must be as much general as possible. It is worth citing Jones and Zhang (1997), Cressie and Huang (1999), Christakos (2000), Brown et al. (2000), de Iaco et al. (2001), de Cesare et al. (2001a, b), Gneiting (2002), Ma (2003a, b), Stein (1999, 2005) and Fernández-Casal (2003). Most of the contributions regard stationary spatio–temporal covariances assuming isotropy on space and time. Particularly, Cressie and Huang (1999) proposed a spectral approach to obtain spatio–temporal covariances, and Gneiting’s (2002) work represent the natural generalisation to this approach, obtained using completely monotone functions and functions whose first derivative is completely monotone. Stein (2005) puts emphasis on the spectral approach. Finally, Fernández-Casal (2003) extends the Shapiro–Botha approach to build flexible variograms in the spatio–temporal context.

Three expressions have been generally used to describe a spatio–temporal random field: a dynamic stochastic representation, a spatio–temporal covariance function, and a spatio–temporal spectral density function (if the spatio–temporal process is stationary).

A stochastic (partial) differential equation could be formally used to describe a spatio–temporal random field, although some of the (partial) derivatives in the equation often do not exist in the usual sense. A deterministic physical model often takes the form of a ordinary or partial differential equation. A dynamic or stochastic model may be formally described by a stochastic (partial) differential equation. Whittle (1954) described a spatial random field by using the stochastic partial differential equation driven by white noise. His work was extended by Heine (1955), Whittle (1956, 1962) and Renshaw (1994). Jones and Zhang (1997) considered a separable and nonseparable spatio–temporal random field defined by two different stochastic partial differential equations giving valid correlation functions. See also Christakos and Hristopulos (1998), and Kolovos et al. (2004) for other models generated from stochastic partial differential equations. There are dynamic space–time models represented using autoregressive time-series models on discrete time points, where Kalman-filter recursions can be applied to obtain optimal predictors. See for example, Huang and Cressie (1996), Mardia et al. (1998), Wikle and Cressie (1999), Stroud et al. (2001), and Huang and Hsu (2004). See also the excellent review on space–time Kalman filtering in Cressie and Wikle (2002).

The spectral density function is an important tool to characterise a stationary spatio–temporal random field, just like the covariance function. Mathematically speaking, the spectral density function and the covariance function are closely related as the pair of the Fourier transform. Thus, the spectral density approach is particularly helpful if there is no explicit expression for the covariance function. Stein (2005) considers a parametric class of spectral densities obtaining some desirable features of the class. In particular, it is noted that it may be desirable to use covariance functions that are smooth away from the origin, and that it can be proved that a covariance function is smooth away from the origin by proving its corresponding spectral density has derivatives with certain moments. These results are used to build covariance functions that are infinitely differentiable away from the origin and allow for arbitrary and possibly different degrees of smoothness for the process in space and time.

A separable spatio–temporal covariance model does provide a useful base for deriving nonseparable spatio–temporal covariance models through appropriate mixture procedures, as shown in Cesare et al. (2001a, b) and Ma (2002, 2003a, b). Ma (2005a) analyses the linear combination of two spatial or spatio–temporal covariances (or variograms) which are isotropic in space. Furthermore, in Ma (2005b) linear combinations of spatio–temporal covariances, where the resulting covariance functions are stationary in time but may not be in space, are presented. New spatio–temporal covariances may be obtained by taking partial derivatives of a differentiable spatio–temporal covariance, whenever it exists. Examples can be found in Ma (2005b).
The theory of spatio-temporal ordinary and generalised random fields also provides a suitable framework for the statistical analysis and modelling of spatio-temporal data (Angulo et al., 2000; Anh and Leonenko, 2001, 2002).

The different methods recalled here and those shown in Kolovos et al. (2004) give an overall picture of the current developments on the construction of spatio-temporal covariance models. However, there has been little direct work examining the consequences of using various types of models. Additionally, in practice, a tradeoff between mathematical tractability of the analytical expressions and incorporation of more realistic features or physical meanings into the model is often required. This paper focuses on stationary models and our attention in this paper is mostly paid to the covariance function, a widely used measure for space and time interaction as well as dependence. The direct use of the covariance function in data analysis include those for evaluating the simple kriging predictor and for calculating the likelihood function.

In practical modelling, the researcher has to face the important decision of which spatio-temporal model is the one that best fits the empirical data. To the best of our knowledge, actually there is no published paper presenting an intensive simulation study comparing the performance of various such models with the aim of helping the researcher in model selection. Thus, the goal of this paper is to compare the spatio-temporal prediction performances among various space–time models, including those having separable spatio-temporal covariance functions and non-separable ones, under several scenarios. Computational performances among various methods will also be compared. In particular the classes of models used here are: (a) the separable model in Huang and Cressie (1996); (b) the class of models given in Cressie and Huang (1999); (c) the class of nonseparable models presented in Gneiting (2002); and (d) a class of nonseparable models based on a simple product–sum structure, as presented in de Cesare et al. (2001a). These classes of models cover quite a portion of spatio–temporal covariance functions used in the literature.

In the space–time modelling literature, model parameters are traditionally estimated by using some method-of-moments estimation tools. In this paper, a more efficient maximum-likelihood method is used. We also address an important issue, which has somewhat been ignored in the past literature, on how to select an appropriate space–time model by accounting for the tradeoff between goodness-of-fit and model complexity. Performances of two commonly used model-selection criteria, Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criterion (BIC) (Schwarz, 1978), are examined.

Note that the asymptotic theory of both AIC and BIC for geostatistical models has not been fully studied in the literature, and thus it is of great interest to investigate whether the traditional consistent result of BIC and the asymptotic efficiency result of AIC in regression and time series model selection would yet hold in the framework of space–time modelling. However, this problem is highly nontrivial, because even the maximum-likelihood estimators of some covariance parameters may not be consistent under infill asymptotics (Zhang, 2004), not to mention asymptotic normality. Rather than proving any asymptotic results, our aim here is to remind the readers the importance of model selection for space–time models, a topic that has somewhat been overlooked, and try to highlight throughout our simulation experiments that both AIC and BIC show some promise in space–time model selection.

Finally, but not least, and on a more applied context, we make use of the findings of the simulation study to present a space–time analysis of the surface shortwave radiation budget (SSRB), which is the amount of energy in the solar region of the electromagnetic spectrum (0.2–4.0 μm) absorbed at the surface. The large pixel size of some satellite sensors for climate studies introduces a number of scientific issues that make relevant, and even necessary, to develop a specific methodology. Without doubt, the large size of the reference pixel makes it necessary to study the change of scale processes in detail in order to be able to compare measurements with different scales (namely point measurements, aircraft observations, remote sensing images with different spatial resolutions), and establish ground-measurement sampling criteria to get different area averages and validate large scale pixels. In this respect, remote sensing poses interesting scientific questions that spatial statistics can help to deal with. In particular, spatial statistics addresses the classification of images by supervised and unsupervised procedures, and the estimation of spatial structures by variograms, interpolation, simulation, sampling and issues of scale. Thus, a space–time analysis can reveal interesting and useful information in, for example, image analysis or remote sensing disciplines.

The plan of the paper is as follows. Section 2 introduces basic concepts for the statistical formulation needed throughout the paper. The various space–time models used in this paper are also described. A comprehensive and intensive simulation study is given in Section 3. Section 4 shows the spatio–temporal analysis of the SSRB data through the models presented in Section 2. The paper ends with some conclusions and discussions.
2. Statistical formulation and methodology

2.1. Setup

Consider a real-valued spatio–temporal random field \( \{Y(s, t), s \in D, t \in \mathbb{Z}\} \) over the spatial domain \( D \subset \mathbb{R}^2 \) and the temporal domain \( \mathbb{Z} \). Its covariance function, \( C(s_1, s_2, t_1, t_2) \), and semi-variogram, \( \gamma(s_1, s_2, t_1, t_2) \), are defined by

\[
C(s_1, s_2, t_1, t_2) = E \left[ \{Y(s_1, t_1) - EY(s_1, t_1)\} \{Y(s_2, t_2) - EY(s_2, t_2)\} \right],
\]

\[
\gamma(s_1, s_2, t_1, t_2) = \frac{1}{2} \text{Var} \left[ \{Y(s_1, t_1) - EY(s_1, t_1)\} - \{Y(s_2, t_2) - EY(s_2, t_2)\} \right],
\]

with \((s_1, t_1), (s_2, t_2) \in D \times \mathbb{Z}\), and whenever they exist.

Under the assumption that \( \text{Var}(Y(s, t)) < \infty \) for all \((s, t) \in D \times \mathbb{Z}\), the covariance function and the semi-variogram are well-defined and related through

\[
\gamma(s_1, s_2, t_1, t_2) = \frac{1}{2} \left[ C(s_1, s_1, t_1, t_1) + C(s_2, s_2, t_2, t_2) \right] - C(s_1, s_2, t_1, t_2).
\]

The random field \( Y(s, t) \) is said to be (weakly, or second-order) stationary in time if \( EY(s, t_1) = EY(s, t_2) \), and \( C(s_1, s_2, t_1, t_2) \) depends only on \( t_1 - t_2 \) for all \( s_1, s_2 \in D \). It is said to be stationary in space if \( EY(s_1, t) = EY(s_2, t) \), and \( C(s_1, s_2, t_1, t_2) \) depends only on \( s_1 - s_2 \) for all \( t_1, t_2 \in \mathbb{Z}\). Finally, the random field is said to be stationary in space–time if \( EY(s_1, t_1) = EY(s_2, t_2) \), and \( C(s_1, s_2, t_1, t_2) \) depends only on \( s_1 - s_2 \) and \( t_1 - t_2 \). When the random field is stationary in time (or space), its covariance is denoted by \( C(s_1, s_2, t) \) (or \( C(s_1, t) \)). When the random field is stationary in space–time, its covariance is denoted by \( C(s, t) \) and called a stationary covariance function. In the same way, the intrinsic stationarity is characterised in terms of the variogram. Equivalently, a constant-mean random field \( Y(s, t) \) is intrinsically stationary in space–time if the increment process \( Y(s + s_0, t + t_0) - Y(s, t) \) is stationary in space–time for any fixed \( (s_0, t_0) \in D \times \mathbb{Z}\). The intrinsic stationarity is more general than the second-order stationarity, since they are processes which are intrinsically stationary but not stationary. If \( Y(s, t) \) is stationary, then it is also intrinsically stationary.

In particular, in this paper we consider \( \{Y(s, t); s \in D \subset \mathbb{R}^2, t \in \mathbb{Z}\} \) a spatio–temporal Gaussian process of interest. Suppose that \( Y(\cdot, \cdot) \) is stationary with \( E(Y(s, t)) \equiv 0 \) (i.e., there is no trend) and spatio–temporal covariance function

\[
C(h, u) \equiv Cov(Y(s, t), Y(s + h, t + u)); \quad s, s + h \in D, \quad u, t \in \mathbb{Z},
\]

where \( h \) is the spatial lag and \( u \) the temporal lag. Also suppose that data \( Z(t) = (Z(s_1, t), \ldots, Z(s_n, t))^T; t = 1, \ldots, T \), are observed at \( n \) locations and \( T \) time points according to the following measurement equation

\[
Z(s, t) = Y(s, t) + \varepsilon(s, t),
\]

where \( \varepsilon(\cdot, \cdot) \) is a white-noise process with variance \( \sigma^2 \).

The optimal linear predictor of \( Y(s, T) \) based on \( Z(1), \ldots, Z(T) \) can be written as

\[
\hat{Y}(s, T) = Cov \left( Y(s, T), Z \right) \text{Var}(Z)^{-1} Z \equiv E(Y(s, T) \mid Z),
\]

where \( Z \equiv (Z(1)^T, \ldots, Z(T)^T)^T \). The prediction method is usually referred to as simple kriging. The corresponding prediction-error variance is

\[
\text{Var}(Y(s, T)) - Cov \left( Y(s, T), Z \right) \text{Var}(Z)^{-1} Cov(Z, Y(s, T)).
\]

In this paper we selected four spatio–temporal models, covering the situations of separability and nonseparability. Our Model I (Huang and Cressie, 1996) is a typical class of separable model, in which the spatial covariance function is arbitrary, and the temporal covariance function has an ARMA\((p, q)\) structure. This class of models is quite large, covering many spatio–temporal covariance structures used in the literature. Moreover, inference and prediction can be implemented fast using the Kalman-filter technique. A systematic way to construct classes of nonseparable models can be found in Cressie and Huang (1999). Model II is a class of models given in Cressie and Huang (1999), while Model III (Gneiting, 2002) is a class of nonseparable models that is constructed using an approach extended from Cressie...
and Huang (1999). Model IV (de Cesare et al., 2001a) shows another way to build nonseparable models, which has a simple product–sum structure.

These four classes of models cover quite a portion of spatio–temporal covariance functions used in the literature. There are other types of models, which may be very complicated and are, somewhat, ad hoc to the problem in hand and thus are not being considered here.

2.2. Dynamic space–time models (Huang and Cressie, 1996)

Consider a spatio–temporal stationary process, \( Y(s, t) \), satisfying the relation

\[
Y(s, t) = \alpha Y(s, t - k) + \zeta(s, t); \quad s \in D, \quad t \in \mathbb{Z},
\]

where \( \zeta(\cdot, \cdot) \) is a spatially stationary and temporally independent Gaussian error process. The spatial covariance function of \( Y(\cdot, t) \) is assumed to belong to the Matérn class (Matérn, 1960)

\[
C_Y(h) = \frac{2^{1-v} \Gamma(v)}{\Gamma(v)} (h \parallel h)^v K_v(h \parallel h),
\]

where \( \phi \equiv (b, \nu, \sigma_y^2) \), \( \sigma_y^2 = \text{Var}(Y(s, t)) \), \( b > 0 \) is the scaling parameter, and \( \nu > 0 \) is the smoothness parameter. Then, the spatio–temporal covariance function of \( Y(\cdot, \cdot) \) is given by

\[
C_Y(h, u|\phi, z) = z^{|n|} C_Y(h),
\]

which defines a separable spatio–temporal covariance function.

The optimal predictor of \( Y(s, t) \) based on data \( Z(1), \ldots, Z(t) \) is given by

\[
\hat{Y}(s, t) \equiv E(Y(s, t)|Z(1), \ldots, Z(t)).
\]

The prediction-error covariance of \( Y(s, t) \) and \( Y(s', t) \) based on data \( Z(1), \ldots, Z(t) \) is given by

\[
P_t(s, s') \equiv \text{Cov} \left( \hat{Y}(s, t) - Y(s, t), \hat{Y}(s', t) - Y(s', t) \right)
= \text{Cov} \left( Y(s, t), Y(s', t) | Z(1), \ldots, Z(t) \right).
\]

Huang and Cressie (1996) showed that \( \hat{Y}(s, t) \) and \( P_t(s, s') \), for any \( s, s' \in D \), can be computed recursively, using the Kalman-filter algorithm, as follows:

\[
\hat{Y}(s, t) = z \hat{Y}(s, t - 1) + Q_t(s)^{\dagger} F_t^{-1} \left( Z(t) - \alpha \hat{Y}(t - 1) \right),
\]

\[
P_t(s, s') = P_{t|t-1}(s, s') + Q_t(s)^{\dagger} F_t^{-1} Q_t(s'),
\]

for \( t = 1, \ldots, T \), where

\[
\hat{Y}(t - 1) \equiv \left( \hat{Y}(s_1, t - 1), \ldots, \hat{Y}(s_n, t - 1) \right)^{\dagger},
\]

\[
P_{t|t-1}(s, s') \equiv \alpha^2 P_{t|t-1}(s, s') + \text{Cov} \left( \zeta(s, t), \zeta(s', t) \right),
\]

\[
Q_t(s) \equiv \left( P_{t|t-1}(s_1, s) \ldots, P_{t|t-1}(s_n, s) \right)^{\dagger},
\]

\[
F_t = \left( P_{t|t-1}(s_i, s_j) \right)_{n \times n} + \sigma_y^2 I_n.
\]

The Kalman-filter algorithm starts with \( \hat{Y}(s, 0) \equiv 0 \) and \( P_{1|0}(s, s') \equiv \text{Cov} \left( Y(s, t), Y(s', t) \right) \).

The joint distribution of \( Z \) can be written as

\[
P(Z) = P(Z(1)) \prod_{t=2}^{T} P(Z(t)|Z(1), \ldots, Z(t - 1)),
\]
where

\[ Z(t)|Z(1), \ldots, Z(t-1) \sim Gau\left(z\hat{Y}(t-1), F_t\right). \]

Consequently, the log-likelihood function of \( \theta = (a, b, \nu, \sigma_Y^2, \sigma_Z^2) \) based on \( Z = (Z(1)^t, \ldots, Z(T)^t)^t \) is given by

\[
L(\theta; Z) = \log(2\pi)^{-nT/2} - \frac{1}{2} \left\{ \sum_{t=1}^{T} \left( \log|F_t| + (Z(t) - z\hat{Y}(t-1))^T F_t^{-1} (Z(t) - z\hat{Y}(t-1)) \right) \right\}.
\]

Note that \( \{F_t\} \) and \( \{\hat{Y}(t)\} \) depend on the parameter vector \( \theta \) and can be computed recursively from the Kalman filter. Therefore, \( L(\theta; Z) \) and the ML estimator of \( \theta \) can be computed efficiently using the Kalman filter. In contrast, Huang and Cressie (1996) used a simple but less efficient method-of-moments estimator of \( \theta \).

2.3. Nonseparable space–time models (Cressie and Huang, 1999)

A five-parameter nonseparable spatio–temporal stationary covariance family given by Cressie and Huang (1999) is of the form

\[
C(h, u|\theta) = \begin{cases} 
\sigma_Y^2 \left( \frac{2\beta}{d/2} \right)^{d/2} \left( \frac{b}{a^2u^2 + 1} \right)^{1/2} \frac{1}{2} \left( \frac{a^2u^2 + 1}{a^2u^2 + \beta} \right)^{1/2} \|h\|^v \\
\times K_v \left( b \left( \frac{a^2u^2 + 1}{a^2u^2 + \beta} \right)^{1/2} \|h\| \right) \\
\sigma_Z^2 \beta^{d/2} \\
\left( \frac{a^2u^2 + 1}{a^2u^2 + \beta} \right)^{d/2}
\end{cases}
\]

if \( \|h\| > 0 \)

where \( \theta = (a, b, \beta, \nu, \sigma_Y^2, \sigma_Z^2)^t \), \( K_v \) is the modified Bessel function of the second kind of order \( \nu \), \( a \geq 0 \) is the scaling parameter of time, \( b \geq 0 \) is the scaling parameter of space, \( \beta > 0 \) is a space–time interaction parameter, \( \nu > 0 \) is a smoothness parameter, and \( \sigma_Y^2 = C(0, 0) > 0 \). Notice that a separable covariance function is obtained when \( \beta = 1 \). Also, in this case, \( C(h, 0|\theta) \) corresponds to the Matérn class.

The log-likelihood function of \( \theta \) based on \( Z = (Z(1)^t, \ldots, Z(T)^t)^t \) can be written as

\[
L(\theta; Z) = \log(2\pi)^{-nT/2} - \frac{1}{2} \left( \log|\Sigma| + Z^T \Sigma^{-1} Z \right),
\]

where \( \Sigma \equiv \text{Var}(Z) \). Note that the \((i, j)\)th entry of \( \Sigma \) is given by

\[
C(s_{i^*} - s_{j^*}, t_i - t_j|\theta) + \sigma_Z^2 \delta_{ij},
\]

where \( i^* \equiv i - n[(i - 1)/n] \), \( j^* \equiv j - n[(j - 1)/n] \), \( t_i \equiv [(i - 1)/n] \), \( t_j \equiv [(j - 1)/n] \), \([x]\) is the largest integer smaller or equal to \( x \), and \( \delta_{ij} \) is the Kronecker \( \delta \) function.

2.4. Nonseparable space–time models (Gneiting, 2002)

A seven-parameter nonseparable spatio–temporal stationary covariance family given by Gneiting (2002) is of the form

\[
C(h, u|\theta) = \begin{cases} 
\sigma_Y^2 \left( \frac{2\beta}{d/2} \right)^{d/2} \left( \frac{b\|h\|}{(a^2u^2 + 1)^{\beta/2}} \right)^v \\
\times K_v \left( \frac{b\|h\|}{(a^2u^2 + 1)^{\beta/2}} \right) \\
\sigma_Z^2 \beta^{d/2} \\
\left( \frac{a^2u^2 + 1}{a^2u^2 + \beta} \right)^{d/2}
\end{cases}
\]

if \( \|h\| > 0 \)

where \( \theta = (a, b, \beta, \nu, \sigma_Y^2, \sigma_Z^2, \sigma_W^2)^t \), \( K_v \) is the modified Bessel function of the second kind of order \( \nu \), \( a \geq 0 \) is the scaling parameter of time, \( b \geq 0 \) is the scaling parameter of space, \( \beta > 0 \) is a space–time interaction parameter, \( \nu > 0 \) is a smoothness parameter, \( \sigma_Y^2 = C(0, 0) > 0 \), and \( \sigma_W^2 \) is the additional variance parameter. Notice that a separable covariance function is obtained when \( \beta = 1 \). Also, in this case, \( C(h, 0|\theta) \) corresponds to the Matérn class.
where $\theta \equiv (a, b, x, v, k_1, b_1, b_2, \beta, v, \delta, \sigma_\tau^2, \sigma_\tau^2)\}^T$, $a > 0$ and $b > 0$ are scaling parameters of time and space, respectively, $x \in (0, 1]$ is the smoothness parameter of time, $v > 0$ is the smoothness parameter of space, $\beta \in [0, 1]$, $\delta \geq 0$, and $\sigma_\tau^2 > 0$. Notice that a separable covariance function is obtained when $\beta = 0$, and $C(h, 0|\theta)$ corresponds to the Matérn class. The log-likelihood function of $\theta$ based on $Z$ is given in the form of (3).

2.5. Product–sum models (de Cesare et al., 2001a)

A seven-parameter nonseparable spatio–temporal stationary covariance family is given by

$$C(h, u|\theta) = \sigma_v^2 (k_1 C_s(h)C_t(u) + k_2 C_s(h) + (1 - k_1 - k_2) C_t(u)),$$

where $\theta \equiv (k_1, k_2, b, x, v, \sigma_v^2, \sigma_\tau^2)\}^T$, $C_s(h) = (2^{1-v}/\Gamma(v))(b\|h\|)^v K_v(b\|h\|)$ is the Matérn spatial correlation class, and $C_t(u) = x^{[u]}$ is the $AR(1)$ autocorrelation function. The log-likelihood function of $\theta$ based on $Z$ is given in the form of (3).

2.6. Parameter estimation and model selection

We used the more efficient maximum likelihood estimation (MLE) rather than the traditional and commonly used method-of-moments. In particular, MLE is based on minimising the minus (log-)likelihood function of the corresponding model,

$$\hat{\theta} = \arg\min_{\theta} -L(\theta; Z).$$

Rather than entering into a fruitless debate over which approach is better, it is more appropriate to address the question through model selection, which makes the answer problem-specific. Methods of cross-validation, minimising an information criterion, or minimising description length offer alternative ways to discriminate between different models of the same data. Little is known about their properties in the spatial context, a topic of research that needs attention in the future.

Thus, concerning model selection, we focus on both, the Akaike information criterion (AIC) (Akaike, 1973)

$$AIC = -2 \log L(\hat{\theta}; Z) + 2|\theta|,$$

and the Bayesian Information Criterion (BIC) (Schwarz, 1978)

$$BIC = -2 \log L(\hat{\theta}; Z) + \log(nT)|\theta|.$$

The properties of AIC and BIC are not well established for space–time models. Recall that, in general, asymptotic efficiency and consistency are viewed as incompatible properties. AIC is the most widely-known representative of the class of asymptotically efficient criteria, and BIC is the most widely-known representative of the class of consistent criteria (McQuarrie and Tsai, 1998).

3. Simulation study

In this section, we describe a thorough simulation study examining and comparing the performance of various spatio–temporal models. Simulations were performed under different scenarios, including separable and nonseparable spatio–temporal covariance functions, various intensities of spatial and temporal dependencies, various degrees of smoothness, and various noise levels. Specifically, we considered the four classes of spatio–temporal models presented in Section 2 with the following combinations of model parameters:

- Model I: $b \in [0.2, 0.5]$, $a \in [0.4, 0.8]$, $v \in [0.5, 1]$, $\sigma_v^2 = 5$, $\sigma_\tau^2 \in \{1.5\}$.
- Model II: $a = 1$, $b \in [0.5, 1.5]$, $v \in [0.5, 1]$, $\sigma_v^2 = 5$, $\sigma_\tau^2 \in \{1, 5\}$.
- Model III: $a = 0.5$, $b \in [0.3, 1.5]$, $a \in [0.1, 0.9]$, $\beta \in [0, 0.9]$, $v \in [0.5, 1]$, $\delta = 0.5$, $\sigma_v^2 = 5$, $\sigma_\tau^2 = 1$.
- Model IV: $(k_1, k_2) \in \{(0.8, 0.1), (0.4, 0.3)\}$, $b \in [0.3, 1.1]$, $a \in [0.4, 0.8]$, $v \in [0.5, 1]$, $\sigma_v^2 = 5$, $\sigma_\tau^2 = 1$. 

With the aim of covering as much practical possibilities as possible, the parameter values were chosen as follows. For Model I, \( b = 0.2 \) (0.5) corresponds to a large (small) spatial range, \( z = 0.4 \) (0.8) corresponds to a small (large) temporal dependence structure, and \( v \) is a smoothness parameter in space with \( v = 1 \) corresponding to a smoother process than \( v = 0.5 \). For Model II, \( b = 1 \) (5) corresponds to a large (small) spatial range, \( \beta \) is the space–time interaction parameter with \( \beta = 1 \) corresponding to a separable spatio–temporal covariance function, and \( v \) is a smoothness parameter with \( v = 1 \) corresponding to a smoother process than \( v = 0.5 \). For Model III, \( b = 0.3 \) (1.5) corresponds to a large (small) spatial range, \( z = 0.3 \) (0.9) corresponds to a small (large) temporal dependence structure, \( \beta \) is the space–time interaction parameter with \( \beta = 0 \) corresponding to a separable spatio–temporal covariance function, and \( v \) is a smoothness parameter with \( v = 1 \) corresponding to a smoother process than \( v = 0.5 \). For Model IV, different values of \( (k_1, k_2) \) correspond to different weights to the mixture of components, \( b = 0.3 \) (1.5) corresponds to a large (small) spatial range, \( z = 0.4 \) (0.8) corresponds to a small (large) temporal dependence structure, and \( v \) is a smoothness parameter in space with \( v = 1 \) corresponding to a smoother process than \( v = 0.5 \).

We generated spatio–temporal variables \( \{Y(s(i, j), t) : i, j = 1, \ldots, 10, \ t = 1, \ldots, 10\} \) on a regular grid of \((10 \times 10) \times 10\) space–time points, where \( s(i, j) \equiv (i - 0.5, j - 0.5)/10 \in [0, 1] \times [0, 1] \). Out of the 100 spatial locations, 20 locations were randomly selected. Then, a random sample was taken at these 20 locations for \( t = 1, \ldots, 10 \) with additive white noise

\[
Z(s(i, j), t) = Y(s(i, j), t) + \varepsilon(s(i, j), t) : \ k = 1, \ldots, 20, \ t = 1, \ldots, 10,
\]

resulting in a dataset of sample size 200, where \( \{\varepsilon(s(i, j), t)\} \) are Gaussian white-noise variables with variance \( \sigma^2_\varepsilon \).

For each dataset, the four classes of spatio–temporal models given in Section 2 were used to fit the data. To speed up the simulation times, and to stabilise the minimisation algorithm, the noise variance \( \sigma^2_\varepsilon \) was assumed known throughout the simulation experiment. Another advantage of assuming \( \sigma^2_\varepsilon \) to be known is that this is a key parameter, and may not be well estimated by MLE.

The prediction was obtained, using the simple-kriging predictors (see Eq. (1)), by plugging in the fitted spatio–temporal covariance structure, and the performances among various methods were compared using the average mean-squared-prediction-error (MSPE) criterion at the last time, i.e. time 10, as follows:

\[
\text{MSPE} = \frac{1}{10^2} \sum_{i=1}^{10} \sum_{j=1}^{10} E\left( \hat{Y}(s(i, j), 10) - Y(s(i, j), 10) \right)^2.
\]

Note that we can take advantage of the fact that the MSPE under the true model can be calculated exactly, and does not need to be simulated. Then, using standard properties of Gaussian distributions, we have that each term \( E\left( \hat{Y}(s(i, j), t) - Y(s(i, j), t) \right)^2 \) appearing in (4) can be decomposed as

\[
E\left( \hat{Y}(s(i, j), t) - Y(s(i, j), t) \right)^2 = E\left( \hat{Y}^*(s(i, j), t) - Y(s(i, j), t) \right)^2 + E\left( \hat{Y}(s(i, j), t) - \hat{Y}^*(s(i, j), t) \right)^2,
\]

where \( \hat{Y}^*(s(i, j), t) \) is the optimal (best linear) predictor of \( Y(s(i, j), t) \) under the true model.

Since the first term on the right-hand side of (5) can be computed exactly using the usual formula for the MSPE of the best linear predictor (see Eqs. (2) and (4)), we only need to approximate the second term by using Monte Carlo simulation. Note that to obtain this second term via simulation, one only needs to simulate \( Y(s(i, j), t) \) at data locations, which clearly speeds up the simulation study. Incidentally, (5) proves that the true model must give smaller MSPE than any fitted model, and this will be confirmed by the simulation results. However, although the MSPE for the true model can be computed theoretically, it depends on the sampling locations. Thus, the MSPE for the true model is reported by averaging over 100 different sampling locations used in the simulations.

It is necessary to stress that even if starting with the optimal initial parameters, we may easily face the problem of convergence to local (but not global) minima, resulting in unusual large MSPE values. To check for convergence, we compared the four likelihood (one for each model) values or the corresponding AIC values. It gives a pretty precise indication of convergence problems if one is significantly larger than the others. In particular, an automatic but effective
method is used here to remove those replicates with outlying MSPE values: a particular replicate is removed when the maximum MSPE values among the four models is 10 times larger than the corresponding minimum for each case.

An important message involved in estimating parameters of these kinds of models, is that the likelihood surface for each of these space–time models may contain many local minima, which is different from the likelihood surface of a purely spatial model using, for example, the Matérn class. Hence, to obtain the ML estimator, it is suggested to start with good several sets of initial parameters and check for goodness-of-fit.

The results are summarised in form of figures (the corresponding tables are available upon request for the interested reader). In particular, Figs. 1–4 show segment-shape plots of mean ± 1.96 × standard error of MSPE values based on 100 replications corresponding to any of the above-mentioned parameter combinations. In particular, the MSPE performance of the various methods when data are generated from each one of the four considered models is analysed.

We considered AIC and BIC criteria to select among the four classes of models. For each replicate, the predicted surface corresponding to AIC (BIC) is equal to the one of the four predicted surfaces having the smallest AIC (BIC) value. The simulation was done on a collection of 16 Linux PCs (CPU: Athlon 1.6 GHz) using the R language.

The degree of smoothing and the noise level (σ^2) defining a space–time process play a crucial role in terms of prediction. In particular, the larger (smaller) the degree of smoothing (noise level) is, the more reliable prediction
is. A separable structure, such as that of Model I, can provide good prediction results even for data generated from nonseparable structures, for example such structures of Models II and IV. In fact (see Figs. 2 and 4), and under many different parameter combinations, Model I provides the best prediction results when data is generated by the nonseparable structures defined by Models II and IV. This behaviour is due to the intrinsic flexibility of Model I, in which the spatial covariance follows a Matérn structure, and the temporal covariance is defined through an ARMA process. Indeed, Model I covers a large number of space–time covariance structures. However, Model I tended to perform worse under stronger spatial dependence.

As expected, see previous comments concerning (5), prediction based on the true model performed better than that based on a fitted model. However, the relative complexity of a class of models prevents prediction based on the correct class of models from providing, in all situations, better results than prediction based on a wrong class of models. For example, the parameters of Models II and IV are more difficult to estimate (compared to the other two models), and a direct consequence is that data initially generated by these two models is not better fitted by themselves.

Both AIC and BIC did as well as the true class of models, in fact AIC and BIC are hardly distinguishable. Maybe BIC performed slightly better than AIC. Note that the asymptotic properties of AIC and BIC have not been quite well
studied for spatial or space–time models, especially when data locations are irregularly located. Some recent results can be found in Hoeting et al. (2006), where, in their conclusions section, it is pointed out the intrinsic difficulty in proving some asymptotic theory of AIC and BIC. If traditional results for regression and time-series hold in this context, then BIC is consistent, but AIC is not (c.f. McQuarrie and Tsai, 1998). This could be one reason why BIC outperforms AIC. However, these asymptotic results have yet to be developed for space–time model selection, in particular, for data that are sampled irregularly in space. Other reasons in favour of BIC are: (a) the sample size considered here is not large enough to satisfy asymptotic conditions; (b) in our setting, one of the four fitted candidate models has the same structure as the true model. Thus, a consistent criterion (BIC) should outperform a criterion which is not consistent (AIC).

It is generally difficult to see clearly the asymptotic properties of AIC and BIC just through a simulation experiment. Basically, AIC (BIC) tends to select a model with larger (smaller) number of parameters by having a smaller penalty. Hence, we see that BIC does better than AIC for Model I, whereas AIC does better than BIC for Model III. In fact, even if BIC is consistent, but AIC is not, we would still see this kind of behaviour unless the sample size is extremely large.

Finally, the underlying spatial process is rarely known in practice. Therefore, it is recommended to fit several different classes of models and use either AIC or BIC for model selection.
### 4. Analysis of the surface shortwave radiation budget (SSRB)

#### 4.1. The data

The SSRB is the amount of energy in the solar region of the electromagnetic spectrum (0.2–4.0 μm) absorbed at the surface. The best way of monitoring these data over extended regions is by means of satellites orbiting the Earth. Due to the fact that satellites provide top of atmosphere (TOA) measurements, the SSRB, measured in W/m², has to be modelled from those measurements, jointly with information about the state of the atmosphere and the surface. A global perspective of the SSRB, both in shortwave and in longwave parts of the spectrum has been achieved during the last years thanks to the SSRB Projects (Whitlock et al., 1995; Stackhouse et al., 1999). However, regional, high spatial resolution description of the SSRB is also of interest as indicator of changes and because of its relationship to aridification from well developed vegetation (Bolle, 1999). Besides, as a measurement of the available energy, SSRB is a key parameter for the derivation of the surface energy fluxes.

The SSRB database has been derived from satellite data at the top of the atmosphere and a physical model that relates the shortwave radiation budget at the top of the atmosphere with that at the surface, making use of ancillary information of precipitable water content in the atmospheric column (Li et al., 1993). For the derivation of the SSRB over the selected study area in the Iberian Peninsula (Fig. 5), we used the Mediterranean Extended Daily One km AVHRR Data.
Fig. 5. Selected region $D$, with latitudes between 37.60°N and 40.09°N and longitudes between 3.92°W and 0.93°W, where the SSRB data set is analysed.

Set (MEDOKADS) (Koslowsky, 1996). The MEDOKADS database is a high spatial resolution database of 10-day maximum NDVI (normalized difference vegetation index) composites, obtained from the AVHRR (advanced very high resolution radiometer) sensors, onboard NOAA (National Oceanic and Atmosphere Administration) satellites. The area consists of land locations of Castilla-La Mancha and the Comunidad Valenciana running southwest to northeast and northwest to southeast, with latitudes between 37.60°N and 40.09°N, and longitudes between 3.92°W and 0.93°W. Due to the high spatial resolution, a square of size 0.01°, the process of SSRB can be considered to be based on point support (and so neglecting any subgrid variability), rather than block support and, consequently, the selected area $D$ configures a rectangular grid of 300 × 250 points. A more detailed description of the derivation of the SSRB dataset can be consulted in Bodas-Salcedo et al. (2003).

With respect to the temporal component, the SSRB data were measured during year 1996. Five measures were taken, each corresponding to a period of 10 days, from April 1 to May 20. Fig. 6 shows the evolution of the stochastic process formed by the SSRB data measured over the 75,000 points (300 × 250) over the five analysed temporal instants.

4.2. Spatio–temporal statistical analysis

Each observation of the SSRB data set is supposed to be a realisation from a stochastic process $Z(s, t)$, which is observed at the spatial locations $s$ and at temporal instants, $t$, corresponding to the period of 10 days, with $s \in D$ and $t \in \mathbb{Z} = \{10, \ldots, 14\}$. With the aim of using the statistical models described in this paper, we decomposed the process $Z(s, t), s \in D, t \in \mathbb{Z}$ into a sum of two components: a trend or large-scale variation $\mu(s, t) = \mu$, representing the mean of the process over the observed region and during the temporal interval considered, and a residual process or small-scale variation $\delta(s, t)$. Thus,

$$Z(s, t) = \mu + \delta(s, t).$$

The trend $\mu$ was estimated as the mean of the observed values, allowing to estimate the residuals by removing them from the observations, $\hat{\delta}(s_1, t) = Z(s, t) - \hat{\mu}$. The resulting stochastic process, $\hat{\delta}(\cdot, \cdot)$, defined the target process over which the four spatio–temporal models used in this paper were applied.

The original region $D$ in Fig. 5 was divided into 30 × 25 subregions of 10 × 10 points each, allowing us to be consistent with the regions used throughout the simulation study. Note that we do not require the stationarity assumption for the SSRB process over the whole region of 300 × 250 points shown in Fig. 6. Since data are eventually analysed locally for each small subregion, it is enough to assume that the SSRB process is stationary over each subregion. In fact, we can even allow different means for different subregions. The SSRB process seems to be spatially nonstationary after looking at Fig. 6. Thus, the data can either be explained by a process with very strong temporal dependence, or perhaps more
appropriately, by a process having heterogeneous means over space. After substracting the mean for each subregion, the process is expected to look less temporally correlated and more spatially stationary.

An exploratory analysis, not reported here, confirmed that the stochastic process \( \hat{Z}(\cdot, \cdot) \) behaved in a similar way in each of these small subregions. As an example, Fig. 7 shows the evolution of the stochastic process \( Z(s, t) \) formed by the SSRB data measured over one of these 30 \( \times \) 25 subregions, randomly chosen from the whole region in \( D \). This selected grid was located between the longitudes \(-3.92\) and \(-3.83\), and the latitudes 38.10 and 38.19. Moreover, Fig. 8 shows the standard deviation against the mean of the process over the five time instants analysed, each corresponding to a period of 10 days, from April 1 to May 20 of 1996.

The four models presented in this paper were used to estimate the SSRB data over each one of the 30 \( \times \) 25 subregions, above mentioned, at the last time \( T = 5 \). To compute the MSPE values in the data analysis, we note the following.
Under the model $Z(s, t) = \mu + Y(s, t) + \varepsilon(s, t)$, where $\mu + Y(s, t)$ is the true SSRB value and $\varepsilon(s, t)$ is the measurement error, we only have the noisy measurements $Z(s, t)$ and the true values are unknown. Hence, we cannot compute $(\hat{Y}(s, t) - Y(s, t))^2$ at the last time $t = T$. Thus, we used a widely known validation method by dividing the data into two parts: one part having sample size $n = 20$ for model fitting, and the other part, with size $80 = N^2 - n$, is used for validation (testing). That is, we computed $\hat{Z}(s, t)$ using the data observed at the $n = 20$ sites, and computed the average of $(\hat{Z}(s, t) - Z(s, t))^2$ over the remaining $80 = N^2 - n$ sites for testing.

A summary, in terms of means and standard errors, of the MSPE results provided by these empirical estimations over 750 repetitions is given in Table 1. This table includes two additional columns for the AIC and BIC models, those showing the lowest values for AIC and BIC in each repetition. Fig. 9 shows segment-shape plots of mean ± 1.96 * standard error of MSPE values for the six models. The MSPE values for the six models were quite similar, but a certain discrimination amongst the four models in favour of Model II is observed. In Table 1 it is also shown the percentage of cases in which
Table 1
Mean and standard error of MSPE values for the six models when fitted to SSRB data over the 750 repetitions

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>Std. error</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>42.5379</td>
<td>0.5772</td>
<td>34.40%</td>
</tr>
<tr>
<td>II</td>
<td>41.5036</td>
<td>0.5532</td>
<td>42.93%</td>
</tr>
<tr>
<td>III</td>
<td>42.1175</td>
<td>0.5719</td>
<td>13.60%</td>
</tr>
<tr>
<td>IV</td>
<td>42.6873</td>
<td>0.5842</td>
<td>9.07%</td>
</tr>
<tr>
<td>AIC</td>
<td>42.4536</td>
<td>0.5782</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>42.4483</td>
<td>0.5782</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 8. Variation of standard deviation against the mean of the process over the five time instants analysed.

Fig. 9. Segment-shape plot of mean $\pm 1.96 \times$ standard error of MSPE values for the estimation of SSRB data at time $T = 5$.

Each fitted model was the best, i.e. showed the lowest MSPE, amongst the 750 repetitions (subregions). Model II was the best in almost 43% of the occasions. Note that the data analysis, similar to the simulation study, did another model comparison via a validation method and this analysis supported the simulation results. In addition, Fig. 10 shows the best model (that model with a smallest MSPE) in each of the 750 grids of $10 \times 10$ points, that partitioned the original region. We used a grey-scale code in which the white one represents Model I (M1) and the black one represents Model IV (M4).

Finally, Fig. 11 shows the estimation of SSRB data obtained for the four models over the $10 \times 10$ points in the randomly selected subregion shown in Fig. 7. For comparison purposes, we also show the real values of the SSRB data for $T = 5$ at the 20 sampling positions.
Fig. 10. Best model in each of the 750 grids of $10 \times 10$ points that partitioned the original region (grey-scale: white = $M_1$ to black = $M_4$).

Fig. 11. Real values, $Z$, and prediction of SSRB data, $\hat{Y}_{1}$ (Model I) to $\hat{Y}_{4}$ (Model IV) in the randomly selected grid of $10 \times 10$ (see details in Fig. 7).
5. Conclusions and discussion

There has been quite a bit of research in recent years on modelling of space–time processes, but very little direct work examining the consequences of using various types of models. We have described a through simulation study comparing four different classes of space–time models in terms of their predictive ability. Model selection is analysed through the use of AIC and BIC values. Thus, we fill a blank in the space–time statistical modelling context.

In practical modelling, we first suggest doing some exploratory data analysis for the data. If the data appear to be spatially and temporally stationary after removing a large-scale trend, the next step would be to check whether the spatio–temporal covariance function is separable. This can be done by simply looking at the empirical spatio–temporal covariances or, more formally, by using a test (e.g. Fuentes, 2003). Choose Model I if separability is appropriate, and consider fitting various spatial and temporal covariance models. Yet, if separability may not be appropriate, consider fitting also nonseparable spatio–temporal models. Finally, use either AIC or BIC to select the final model.

Based on our findings, we highlight the following points:

1. The degree of smoothing and the noise level \( \sigma^2 \) defining a space–time process play a crucial role in terms of prediction. In particular, matching a tentative theoretical model with a model for fitting can be less tricky for processes with a larger degree of smoothing and/or with a small noise level.

2. A separable structure can provide good prediction results even for data generated from nonseparable structures. This can be simply due to the intrinsic flexibility of the separable model proposed. For example, a Matérn structure for the spatial covariance, and an ARMA structure for the temporal interactions, can well do this job.

3. Though overstated, it is important to stress that prediction based on the true model performs better than that based on a fitted model.

4. However, the relative complexity of a class of models prevents prediction based on the correct class of models from providing, in all situations, better results than prediction based on a wrong class of models.

5. Both AIC and BIC did as well as the true class of models, in fact AIC and BIC are hardly distinguishable. Maybe, BIC performed slightly better than AIC. This is probably due to the fact that the traditional consistent result of BIC in regression and time series model selection may still hold in space–time model selection considered here. However, as mentioned in Hoeting et al. (2006), even asymptotic theory of the MLE for geostatistical models has not been fully developed when data are irregularly sampled in space. Further investigation is needed to understand the asymptotic properties of AIC and BIC, but this is beyond the scope of this paper.

6. The underlying spatial process is rarely known in practice. Thus, it is usually advised trying various models after an appropriate exploratory data analysis, and doing some diagnostic checks afterwards or using either AIC or BIC for model selection.

7. In terms of computation and the ability to handle large space–time data sets, a separable and flexible model, such as our Model I, would be the choice. However, if nonseparability is guessed, then a flexible, easy-to-compute and interpret nonseparable structure, such as our Model II, should be clearly proposed.

8. Model II showed the best behaviour amongst the four models when the solar radiation data was analysed. This model presented the lowest values of both MSPE and corresponding standard errors. However, the separable structure showed the second best results. So, although the SSRB data evolves jointly in space and time, in a nonseparable way, we could also use other sufficiently flexible separable structures.

Acknowledgements

The Associate Editor and two anonymous referees are acknowledged with thanks. Their precise comments and suggestions have clearly improved an earlier version of the manuscript.

References


