

ADAPTIVELY VARYING COEFFICIENT SPATIO-TEMPORAL MODELS*

Zudi Lu^{1,2,3} Dag Johan Steinskog^{4,5} Dag Tjøstheim⁶ Qiwei Yao^{1,7}

¹Department of Statistics, London School of Economics, London, WC2A 2AE, UK

²Department of Mathematics & Statistics, Curtin University of Technology, Perth, Australia

³ School of Mathematical Sciences, The University of Adelaide, Adelaide, Australia

⁴ Nansen Environmental and Remote Sensing Center, 5006 Bergen, Norway

⁵Bjerknes Centre for Climate Research, 5007 Bergen, Norway

⁶Department of Mathematics, University of Bergen, 5007 Bergen, Norway

⁷Guanghua School of Management, Peking University, Beijing 100871, China

Abstract

We propose an adaptive varying-coefficient spatio-temporal model for data observed irregularly over space and regularly in time. The model is capable of catching possible nonlinearity (both in space and in time) and nonstationarity (in space) by allowing the autoregressive coefficients to vary with both spatial location and an unknown index variable. We suggest a two-step procedure to estimate both the coefficient functions and the index variable, which is readily implemented and can be computed even for large spatio-temporal data sets. Our theoretical results indicate that in the presence of the so-called nugget effect, the errors in the estimation may be reduced via the spatial smoothing – the second step in the proposed estimation procedure. The simulation results reinforce this finding. As an illustration, we apply the methodology to a data set of sea level pressure in the North Sea.

Key words: β -mixing, kernel smoothing, local linear regression, nugget effect, spatial smoothing, unilateral order.

*Partially supported by a Leverhulme Trust research grant and EPSRC research grant. Lu's work was also partially supported by an Internal Research Grant from Curtin University of Technology and the Discovery Grant from Australian Research Council.

1 Introduction

The wide availability of data observed over time and space, in particular through inexpensive geographical information systems, has stimulated many studies in a variety of disciplines such as environmental science, epidemiology, political science, demography, economics and geography. In these studies, the geographical areas (e.g. counties, census tracts) are taken as units of analysis, and specific statistical methods have been developed to deal with the spatial structure reflected in the distribution of the dependent variable; see e.g. Ripley (1981), Anselin (1988), Cressie (1993), Guyon (1995), Stein (1999) and Diggle (2003) for systematic reviews of these and related topics.

In this paper we are concerned with spatio-temporal data which are measured or transformed to a continuous scale and observed irregularly over space but regularly over time. Data sets of this form exist extensively. It may be environmental monitoring stations located irregularly in some region, but with measurements taken daily; see for instance Fotheringham *et al.* (1998), Brunson *et al.* (2001), Fuentes (2001), Zhang *et al.* (2003). Applications in spatial disease modelling may be found in Knorr-Held (2000), Lagazio *et al.* (2001).

Our model is of the form

$$Y_t(\mathbf{s}) = a \{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^\tau \mathbf{X}_t(\mathbf{s})\} + \mathbf{b}_1 \{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^\tau \mathbf{X}_t(\mathbf{s})\}^\tau \mathbf{X}_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}), \quad (1.1)$$

where $Y_t(\mathbf{s})$ is the spatio-temporal variable of interest, and t is time, $\mathbf{s} = (u, v) \in \mathcal{S} \subset \mathbb{R}^2$ is a spatial location. Moreover, $a(\mathbf{s}, z)$ and $\mathbf{b}_1(\mathbf{s}, z)$ are unknown scalar and $d \times 1$ functions, $\boldsymbol{\alpha}(\mathbf{s})$ is an unknown $d \times 1$ index vector, $\{\varepsilon_t(\mathbf{s})\}$ is a noise process which, for each fixed \mathbf{s} , forms a sequence of *i.i.d.* random variables over time, and $\mathbf{X}_t(\mathbf{s}) = \{X_{t1}(\mathbf{s}), \dots, X_{td}(\mathbf{s})\}^\tau$ consists of time-lagged values of $Y_t(\cdot)$ in a neighbourhood of \mathbf{s} and, possibly, some exogenous variables. Throughout the paper we use τ to denote the transpose of a vector or a matrix. We let both the regression coefficient \mathbf{b}_1 and the intercept a depend on the location \mathbf{s} , as well as on the index variable $\boldsymbol{\alpha}(\mathbf{s})^\tau \mathbf{X}_t(\mathbf{s})$, to catch possible nonstationary features over space. Model (1.1) is unilateral in time and, therefore, it can be easily simulated in a Monte Carlo study. Furthermore it is readily applicable to model practical problems. Those features make the model radically different from purely spatial autoregressive models (Yao and Brockwell 2006). Also note that at a given location \mathbf{s} , the coefficient functions are univariate functions of $\boldsymbol{\alpha}(\mathbf{s})^\tau \mathbf{X}_t(\mathbf{s})$. Therefore only one-dimensional smoothing is required in estimating those coefficients. (See Section 2.2.1 below).

Model (1.1) is proposed to model nonlinear and nonstationary spatial data in a flexible manner. Note that the literature on spatial processes is overwhelmingly dominated by linear models. On

the other hand, many practical data exhibit clearly nonlinear and nonstationary features. For example, in our analysis of the daily mean sea level pressure (MSLP) readings in Section 5, the original time series at each location is non-linear. Figure 3 below displays the plots for the MSLP data at five randomly selected locations. There are roughly three high pressure periods within the time span concerned. In the middle column of Figure 3, the correlation structure of the differenced series in the high pressure periods is different from that of the low pressure periods and the transition periods. This naturally calls for a nonlinear model with varying coefficients depending also on the past values of the variables from the neighbourhood locations. Such a dynamics cannot result in a Gaussian marginal distribution, as indicated in the plots on the right column in Figure 3. The estimated densities are more peaked than the normal distribution, the accumulation of density mass around zero being due to the high pressure activity.

Model (1.1) is a spatial extension of the adaptive varying-coefficient linear model of Fan *et al.* (2003), see also Xia and Li (1999). Due to the spatial non-stationarity of the model, we adopt an estimation strategy involving two steps, which also facilitates the *parallel computation* over different locations. First, for each fixed location \mathbf{s} , we estimate the varying coefficient functions a and \mathbf{b}_1 and the index vector $\boldsymbol{\alpha}$ based on the observations at the location \mathbf{s} only. This is a time series estimation problem. Our estimation is based on local linear regression. The second step is to apply spatial smoothing to pool together the information from neighbourhood locations. Asymptotic properties of our estimators have been established, which indicate that in the presence of the so-called ‘nugget effect’, the spatial smoothing will reduce the estimation errors. (See also Lu *et al.* 2008). Our simulation study reinforces this finding.

As far as we are aware, this is the first paper to address spatio-temporal nonlinear dependence structures with spatial nonstationarity and non-Gaussian distributions. Earlier work on nonlinear and/or non-Gaussian spatial models includes, among others, Matheron (1976), Rivoirard (1994), Chilès and Delfiner (1999, Chapter 6), Hallin *et al.* (2004a, 2004b), Biau and Cadre (2004), Gao *et al.* (2006), Lu *et al.* (2007).

The structure of the paper is as follows. We introduce our methodology in Section 2. Illustration with simulated examples are reported in Section 3. In Section 5 there is an application to a data set of sea level pressure in the North Sea, for which the index variable may be viewed as an analogue of a spatial principal component used in the so-called EOF analysis for climate time series. The asymptotic properties are reported in Section 4, with the regularity conditions deferred to an appendix.

2 Methodology

2.1 Identification

Model (1.1) is not identifiable. In fact, it may be equivalently expressed as

$$Y_t(\mathbf{s}) = [a\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^\top \mathbf{X}_t(\mathbf{s})\} + \kappa\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^\top \mathbf{X}_t(\mathbf{s})\} \boldsymbol{\alpha}(\mathbf{s})^\top \mathbf{X}_t(\mathbf{s})] \\ + [\mathbf{b}_1\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^\top \mathbf{X}_t(\mathbf{s})\} - \kappa\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^\top \mathbf{X}_t(\mathbf{s})\} \boldsymbol{\alpha}(\mathbf{s})]^\top \mathbf{X}_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}), \quad (2.1)$$

for any real-valued function $\kappa(\cdot)$. To overcome this problem, we may assume that the last component of $\boldsymbol{\alpha}(\mathbf{s})$ is non-zero, the first non-zero component of $\boldsymbol{\alpha}(\mathbf{s})$ is positive, and $\|\boldsymbol{\alpha}(\mathbf{s})\|^2 = 1$. Based on those assumptions, Fan *et al.* (2003) impose the condition that the last component of \mathbf{b}_1 is 0 in (1.1). Put $\mathbf{b}_1 = (\mathbf{b}^\top, 0)^\top$. This effectively reduces (1.1) to the form

$$Y_t(\mathbf{s}) = a\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^\top \mathbf{X}_t(\mathbf{s})\} + [\mathbf{b}\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^\top \mathbf{X}_t(\mathbf{s})\}]^\top \check{\mathbf{X}}_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}), \quad (2.2)$$

where $\check{\mathbf{X}}_t(\mathbf{s}) = \{X_{t,1}(\mathbf{s}) \cdots, X_{t,d-1}(\mathbf{s})\}^\top$. In fact, $\boldsymbol{\alpha}$, a and \mathbf{b} in (2.2) are now identifiable as long as the regression function $E\{Y_t(\mathbf{s}) | \mathbf{X}_t(\mathbf{s}) = \mathbf{x}\}$ is not of the form $\boldsymbol{\alpha}^\top \mathbf{x} \boldsymbol{\beta}^\top \mathbf{x} + \boldsymbol{\gamma}^\top \mathbf{x} + c$, where $\boldsymbol{\beta}$, $\boldsymbol{\gamma} \in \mathbb{R}^d$, $c \in \mathbb{R}^1$ are constants only related to \mathbf{s} ; see Theorem 1(b) of Fan *et al.* (2003). Note that (2.1) reduces to (2.2) by setting $\kappa(\mathbf{s}, z) = b_{1d}/\alpha_d$, where b_{1j} and α_j denote, respectively, the j -th component of $\mathbf{b}_1(\mathbf{s}, z)$ and $\boldsymbol{\alpha}(\mathbf{s})$.

A disadvantage of the form (2.2) is that all the components of $\mathbf{X}_t(\mathbf{s})$ are not on an equal footing, which may cause difficulties in model interpretation. This may be particularly problematic when $\mathbf{X}_t(\mathbf{s})$ consists of neighbour variables of $Y_t(\mathbf{s})$. One possible remedy is to impose an orthogonal constraint in the model instead, that is

$$Y_t(\mathbf{s}) = a^*\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^\top \mathbf{X}_t(\mathbf{s})\} + \mathbf{b}_1\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^\top \mathbf{X}_t(\mathbf{s})\}^\top \mathbf{X}_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}), \quad \boldsymbol{\alpha}(\mathbf{s})^\top \mathbf{b}_1(\mathbf{s}, z) = 0. \quad (2.3)$$

In fact such an orthogonal representation may be obtained from (2.2) with

$$a^*(\mathbf{s}, z) = a(\mathbf{s}, z) + \boldsymbol{\alpha}(\mathbf{s})(\mathbf{b}(\mathbf{s}, z)^\top, 0)z \quad \text{and} \quad \mathbf{b}_1(\mathbf{s}, z) = (\mathbf{b}(\mathbf{s}, z)^\top, 0)^\top - \boldsymbol{\alpha}(\mathbf{s})(\mathbf{b}(\mathbf{s}, z)^\top, 0)\boldsymbol{\alpha}(\mathbf{s}) \quad (2.4)$$

while $\boldsymbol{\alpha}(\mathbf{s})$ remains unchanged. Note the condition $\boldsymbol{\alpha}(\mathbf{s})^\top \mathbf{b}_1(\mathbf{s}, z) = 0$ is fulfilled as $\|\boldsymbol{\alpha}(\mathbf{s})\|^2 = 1$. Furthermore, (2.3) is an equivalent representation of (2.2), and it is identifiable as long as (2.2) is identifiable. To see this, we first note that (2.2) may be deduced from (2.3) by letting

$$a(\mathbf{s}, z) = a^*(\mathbf{s}, z) + \frac{b_{1d}}{\alpha_d} z \quad \text{and} \quad \mathbf{b}(\mathbf{s}, z) = (b_{11}, \dots, b_{1,d-1})^\top - \frac{b_{1d}}{\alpha_d} (\alpha_1, \dots, \alpha_{d-1})^\top. \quad (2.5)$$

This is effectively to write the first equation in (2.3) in the form of (2.1) with (a, κ) replaced by $(a^*, b_{1d}/\alpha_d)$. Moreover, as $\boldsymbol{\alpha}(\mathbf{s})^\tau \mathbf{b}_1(\mathbf{s}, z) = 0$ and $\|\boldsymbol{\alpha}(\mathbf{s})\|^2 = 1$, it follows from the second equality in (2.5) that

$$(\mathbf{b}(\mathbf{s}, z)^\tau, 0)\boldsymbol{\alpha}(\mathbf{s}) = \sum_{j=1}^{d-1} \alpha_j b_{1j} - \frac{b_{1d}}{\alpha_d} \sum_{j=1}^{d-1} \alpha_j^2 = -\alpha_d b_{1d} - \frac{b_{1d}}{\alpha_d} (1 - \alpha_d^2) = -\frac{b_{1d}}{\alpha_d}.$$

Hence the d -component b_{1d} of $\mathbf{b}_1(\mathbf{s}, z)$ is identifiable when model (2.2) is identifiable. This, together with (2.5), implies that all the other components of $\mathbf{b}_1(\mathbf{s}, z)$ as well as $a^*(\mathbf{s}, z)$ are also identifiable.

Since the orthogonal constraint $\boldsymbol{\alpha}^\tau \mathbf{b}_1 = 0$ in (2.3) poses further technical complication in inference, in the sequel we proceed with the identifiable model (2.2), assuming that the first non-zero element of $\boldsymbol{\alpha}(\mathbf{s})$ is positive. Only for the real data example in section 5, we present the fitted model in the form (2.3) via the transformation (2.4).

2.2 Estimation

With the observations $\{Y_t(\mathbf{s}_i), \mathbf{X}_t(\mathbf{s}_i)\}$ for $t = 1, \dots, T$ and $i = 1, \dots, N$, we estimate a, \mathbf{b} and $\boldsymbol{\alpha}$ in model (2.2) in two steps: (i) For each fixed $\mathbf{s} = \mathbf{s}_i$, we estimate them using the time series data $\{Y_t(\mathbf{s}), \mathbf{X}_t(\mathbf{s}), t = 1, \dots, T\}$ only. (ii) By pooling the information from neighbour locations via spatial smoothing, we improve the estimators obtained in (i). The second step also facilitates the estimation at a location \mathbf{s} with no direct observations (i.e. $\mathbf{s} \neq \mathbf{s}_i$ for $1 \leq i \leq N$).

2.2.1 Time Series Estimation

For a fixed \mathbf{s} , we estimate the direction $\boldsymbol{\alpha}(\mathbf{s})$ and the coefficient functions $a(\mathbf{s}, \cdot)$ and $\mathbf{b}(\mathbf{s}, \cdot)$. Fan *et al.* (2003) proposed a backfitting algorithm to solve this estimation problem. We argue that with modern computer power, the problem can be dealt with in a more direct manner even for moderately large d such as d between 10 and 20. Note that once the direction $\boldsymbol{\alpha}(\mathbf{s})$ is fixed, the estimators for a and \mathbf{b} may be obtained by one-dimensional smoothing over $\boldsymbol{\alpha}(\mathbf{s})^\tau \mathbf{X}_t(\mathbf{s})$ using, for example, the standard kernel regression methods.

We consider the estimation for $\boldsymbol{\alpha}(\mathbf{s})$ first. If the coefficient functions a and \mathbf{b} were known, we may estimate $\boldsymbol{\alpha}$ by solving a nonlinear least squares problem; see (2.7) below. Since a and \mathbf{b} are unknown, we may plug in their appropriate estimators instead. By doing so, we will have used the same observations twice. Note that those estimators themselves are functions of $\boldsymbol{\alpha}$. Hence a cross-validation approach will be employed to mitigate the effect of the double use of the same

data set. For any $\boldsymbol{\alpha}(\mathbf{s})$, define the leave-one-out estimators $\check{a}_{-t}(\mathbf{s}, z) = \check{a}$ and $\check{\mathbf{b}}_{-t}(\mathbf{s}, z) = \check{\mathbf{b}}$, where $(\check{a}, \check{\mathbf{b}})$ minimises

$$\frac{1}{T-1} \sum_{\substack{1 \leq j \leq T \\ j \neq t}} \{Y_j(\mathbf{s}) - a - \mathbf{b}^\tau \check{\mathbf{X}}_j(\mathbf{s})\}^2 K_h\{\boldsymbol{\alpha}^\tau(\mathbf{s}) \mathbf{X}_j(\mathbf{s}) - z\}, \quad (2.6)$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function and $h > 0$ is a bandwidth. Then we choose $\hat{\boldsymbol{\alpha}}(\mathbf{s})$ and h which minimise

$$R(\boldsymbol{\alpha}, h) = \frac{1}{T} \sum_{t=1}^T [Y_t(\mathbf{s}) - \check{a}_{-t}\{\mathbf{s}, \boldsymbol{\alpha}^\tau \mathbf{X}_t(\mathbf{s})\} - \check{\mathbf{b}}_{-t}\{\mathbf{s}, \boldsymbol{\alpha}^\tau \mathbf{X}_t(\mathbf{s})\}^\tau \check{\mathbf{X}}_t(\mathbf{s})]^2 w\{\boldsymbol{\alpha}^\tau \mathbf{X}_t(\mathbf{s})\}, \quad (2.7)$$

where $w(\cdot)$ is a weight function which controls the boundary effect. In the above definition for the estimator $\hat{\boldsymbol{\alpha}}(\mathbf{s})$, we used the Nadaraya-Watson estimators \check{a} and $\check{\mathbf{b}}$ in (2.6) to keep the function $R(\boldsymbol{\alpha}, h)$ simple. The asymptotic property of the estimator $\hat{\boldsymbol{\alpha}}(\mathbf{s})$ has been derived based on the results of Lu *et al.* (2007); see (4.11) in section 4 below. The minimisation of $R(\cdot)$ may be carried out using, for example, the downhill simplex method (section 10.4 of Press *et al.* 1992).

Once $\boldsymbol{\alpha}(\mathbf{s}) = \hat{\boldsymbol{\alpha}}(\mathbf{s})$ is known, we may estimate a and \mathbf{b} using univariate local linear regression estimation. To this end, let $(\hat{a}, \hat{\mathbf{b}}, \hat{c}, \hat{\mathbf{d}})$ be the minimisers of

$$\frac{1}{T} \sum_{j=1}^T [Y_j(\mathbf{s}) - a - c(Z_j - z) - \{\mathbf{b} - \mathbf{d}(Z_j - z)\}^\tau \check{\mathbf{X}}_j(\mathbf{s})]^2 K_{\bar{h}}(Z_j - z), \quad (2.8)$$

where $Z_t = \hat{\boldsymbol{\alpha}}(\mathbf{s})^\tau \mathbf{X}_t(\mathbf{s})$, and a different bandwidth \bar{h} from the h in (2.6) will be used. Note that both h and \bar{h} are depending on T . The estimators for the functions $a(\mathbf{s}, \cdot)$ and $\mathbf{b}(\mathbf{s}, \cdot)$ at z are defined as $\hat{a}(\mathbf{s}, z) = \hat{a}$ and $\hat{\mathbf{b}}(\mathbf{s}, z) = \hat{\mathbf{b}}$.

2.2.2 Spatial smoothing

Although we do not assume stationarity over space, improvement in estimation of a , \mathbf{b} and $\boldsymbol{\alpha}$ over \mathbf{s} may be gained by extracting information from neighbourhood locations, due to the continuity of the functions concerned. Zhang *et al.* (2003) showed for the model they considered that it was indeed the case in the presence of the *nugget effect* (Cressie, 1993, § 2.3.1). See also Lu *et al.* (2008). Furthermore, the spatial smoothing provides a way to extrapolate our estimators to the locations at which observations are not available.

Let $W(\cdot)$ be a kernel function defined on \mathbb{R}^2 . Put $W_{\tilde{h}}(\mathbf{s}) = \tilde{h}^{-2}W(\mathbf{s}/\tilde{h})$, where $\tilde{h} > 0$ is a bandwidth depending on the size N of spatial samples. The bandwidth \tilde{h} is different from both

h and \bar{h} in the last subsection. Based on the estimators obtained above for each of the locations $\mathbf{s}_1, \dots, \mathbf{s}_N$, the spatially smoothed estimators at the location \mathbf{s}_0 are defined as

$$\tilde{\boldsymbol{\alpha}}(\mathbf{s}_0) = \sum_{j=1}^N \hat{\boldsymbol{\alpha}}(\mathbf{s}_j) W_{\bar{h}}(\mathbf{s}_j - \mathbf{s}_0) / \sum_{j=1}^N W_{\bar{h}}(\mathbf{s}_j - \mathbf{s}_0), \quad (2.9)$$

$$\tilde{a}(\mathbf{s}_0, z) = \sum_{j=1}^N \hat{a}(\mathbf{s}_j, z) W_{\bar{h}}(\mathbf{s}_j - \mathbf{s}_0) / \sum_{j=1}^N W_{\bar{h}}(\mathbf{s}_j - \mathbf{s}_0), \quad (2.10)$$

$$\tilde{\mathbf{b}}(\mathbf{s}_0, z) = \sum_{j=1}^N \hat{\mathbf{b}}(\mathbf{s}_j, z) W_{\bar{h}}(\mathbf{s}_j - \mathbf{s}_0) / \sum_{j=1}^N W_{\bar{h}}(\mathbf{s}_j - \mathbf{s}_0). \quad (2.11)$$

2.3 Bandwidth selection

Fan *et al.* (2003) outlined an empirical rule for selecting a bandwidth in fitting non-spatial varying-coefficient regression models. Below we adapt that idea to determining bandwidths \bar{h} in (2.8) and \tilde{h} in (2.9) – (2.11). Note that the bandwidth h in (2.7) is selected, together with $\hat{\boldsymbol{\alpha}}(\mathbf{s})$, by cross-validation.

We first deal with the selection of \bar{h} in (2.8). Let

$$CV_1(\bar{h}) = \frac{1}{T} \sum_{t=1}^T [Y_t(\mathbf{s}) - \check{a}_{-t, \bar{h}}\{\mathbf{s}, \hat{\boldsymbol{\alpha}}^T \mathbf{X}_t(\mathbf{s})\} - \check{\mathbf{b}}_{-t, \bar{h}}\{\mathbf{s}, \hat{\boldsymbol{\alpha}}^T \mathbf{X}_t(\mathbf{s})\}^T \check{\mathbf{X}}_t(\mathbf{s})]^2 w\{\hat{\boldsymbol{\alpha}}^T \mathbf{X}_t(\mathbf{s})\}, \quad (2.12)$$

where $\check{a}_{-t, \bar{h}}\{\mathbf{s}, z\}$ and $\check{\mathbf{b}}_{-t, \bar{h}}\{\mathbf{s}, z\}$ are the leave-one-out estimators defined as in (2.8) but with the term of $j = t$ removed from the sum. Under appropriate regularity conditions (c.f. Lu, Tjøstheim and Yao, 2007), it holds that

$$CV_1(\bar{h}) = c_0 + c_1 \bar{h}^4 + \frac{c_2}{T\bar{h}} + o_P(\bar{h}^4 + T^{-1}\bar{h}^{-1}).$$

Thus, up to first order asymptotics, the optimal bandwidth is $\bar{h}_{opt} = (c_2/4Tc_1)^{1/5}$. In practice, the coefficients c_0 , c_1 and c_2 will be estimated from solving the least squares problem

$$\min_{c_0, c_1, c_2} \sum_{k=1}^q \{CV_{1k} - c_0 - c_1 \bar{h}_k^4 - c_2/(T\bar{h}_k)\}^2, \quad (2.13)$$

where $CV_{1k} = CV_1(\bar{h}_k)$ obtained from (2.12), and $\bar{h}_1, \dots, \bar{h}_q$ are q prescribed bandwidth values. We let $\bar{h} = (\hat{c}_2/4T\hat{c}_1)^{1/5}$ when both \hat{c}_1 and \hat{c}_2 are positive. In the likely event that one of them is nonpositive, we let $\bar{h} = \arg \min_{1 \leq k \leq q} CV_1(\bar{h}_k)$. This bandwidth selection procedure is computationally efficient as q is moderately small, i.e. we only need to compute q CV-values; see Remark 2(c) in Fan *et al.* (2003). Furthermore the least squares estimation (2.13) also serves as a smoother for the CV bandwidth estimates. Also see Ruppert (1997).

The bandwidth \tilde{h} in (2.9) – (2.11) may be determined in the same manner. For example, for the estimator (2.9), the CV function is defined as

$$CV_2(\tilde{h}) = \frac{1}{N} \sum_{i=1}^N [\hat{\alpha}(\mathbf{s}_i) - \tilde{\alpha}_{-\mathbf{s}_i, \tilde{h}}(\mathbf{s}_i)]^2,$$

which admits the asymptotic expansion

$$CV_2(\tilde{h}) = d_0 + d_1 \tilde{h}^4 + \frac{d_2}{N \tilde{h}^2} + o_P(\tilde{h}^4 + N^{-1} \tilde{h}^{-2}).$$

The resulting bandwidth is of the form $\tilde{h} = \{\hat{d}_2 / (2N \hat{d}_1)\}^{1/6}$.

The CV bandwidth selection has been extensively studied in the literature. In the sense of mean integrated squared error (MISE) the relative error of a CV-bandwidth is higher than that of, for example, a plug-in method of Ruppert *et al.* (1995). However there is a growing body of opinion (e.g., Mammen, 1990; Jones, 1991; Härdle & Vieu, 1992; and Loader, 1999) maintaining that the selection of bandwidth should be targeted at estimating the unknown function instead of the ideal bandwidth itself. Therefore, one should seek a bandwidth minimising the integrated squared error (ISE) rather than the MISE, i.e., focusing on loss rather than risk. From this point of view, the CV-bandwidth performs reasonably well (Hall and Johnstone, 1992, page 479). In time series context, the argument for why cross-validation is an appropriate selection method for the bandwidth can be found in Kim and Cox (1996), Quintela-del-Río (1996) and Xia and Li (2002), among others.

3 A simulated example

Consider a spatio-temporal process

$$Y_{t+1}(\mathbf{s}_{ij}) = a\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\} + b_1\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\}Y_t(\mathbf{s}_{ij}) + b_2\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\}X_t(\mathbf{s}_{ij}) + 0.1\varepsilon_t(\mathbf{s}_{ij})$$

defined on the grid points $\mathbf{s}_{ij} = (u_i, v_j)$, where

$$a(\mathbf{s}_{ij}, z) = 3 \exp\left\{\frac{-2z^2}{1 + 7(u_i + v_j)}\right\}, \quad b_1(\mathbf{s}_{ij}, z) = 0.7 \sin\{7\pi(u_i + v_j)\}, \quad b_2(\mathbf{s}_{ij}, z) = 0.8z,$$

$$Z_t(\mathbf{s}_{ij}) = \frac{1}{3}\{2Y_t(\mathbf{s}_{ij}) + 2X_t(\mathbf{s}_{ij}) + X_t(\mathbf{s}_{i,j+1})\}, \quad X_t(\mathbf{s}_{ij}) = \frac{2}{9} \sum_{k=-1}^1 \sum_{\ell=-1}^1 e_t(\mathbf{s}_{i+k, j+\ell}),$$

and all $\varepsilon_t(\mathbf{s}_{ij})$ and $e_t(\mathbf{s}_{ij})$ are independent $N(0, 1)$ random variables. Observations were taken over $N = 64$ grid points $\{(u_i, v_j), 1 \leq i, j \leq 8\}$, where $u_i = v_i = (i - 1)/8$ with $T = 60$ or

100. For each given $\mathbf{s} = \mathbf{s}_{ij}$, we estimated the curves $a(\mathbf{s}, \cdot)$, $b_1(\mathbf{s}, \cdot)$ and $b_2(\mathbf{s}, \cdot)$ on the 11 grid points $z_\ell = -0.5 + 0.1(\ell - 1)$ ($\ell = 1, \dots, 11$), as well as the index $\boldsymbol{\alpha}(\mathbf{s}) = (\alpha_1(\mathbf{s}), \alpha_2(\mathbf{s}), \alpha_3(\mathbf{s}))^\tau \equiv (2/3, 2/3, 1/3)^\tau$, which is independent of \mathbf{s} . The accuracy of the estimation is measured by the the squared estimation errors:

$$\begin{aligned} \text{SEE}\{\widehat{\alpha}_j(\mathbf{s})\} &= \{\widehat{\alpha}_j(\mathbf{s}) - \alpha_j\}^2, \quad j = 1, 2, 3, \\ \text{SEE}\{\widehat{a}(\mathbf{s})\} &= \frac{1}{11} \sum_{\ell=1}^{11} \{\widehat{a}(\mathbf{s}, z_\ell) - a(\mathbf{s}, z_\ell)\}^2, \\ \text{SEE}\{\widehat{b}_k(\mathbf{s})\} &= \frac{1}{11} \sum_{\ell=1}^{11} \{\widehat{b}_k(\mathbf{s}, z_\ell) - b_k(\mathbf{s}, z_\ell)\}^2, \quad k = 1, 2. \end{aligned}$$

For each setting, we replicated the experiments 10 times. The SEE over the 64 grid points are collectively displayed as boxplots in Figure 1 for $T = 60$, and in Figure 2 for $T = 100$ (i.e. each of the boxplots is based on $10 \times 64 = 640$ SEE values). It is clear that the estimates defined in section 2.2.1 are less accurate than those defined in section 2.2.2. In fact the gain from the spatial smoothing is substantial. The plots also indicate that the estimation becomes more accurate when T increases.

The improvement from the spatial smoothing is due to the fact that the functions to be estimated are continuous in \mathbf{s} . If we view the observations $\{Y_t(\mathbf{s}_{ij}), X_t(\mathbf{s}_{ij})\}$ as a sample from a spatio-temporal process with \mathbf{s} varying continuously over space, the sample paths of such an underlying process are discontinuous over space, as both $\varepsilon_t(\mathbf{s})$ and $e_t(\mathbf{s})$ are independent at different locations no matter how close they are. The spatial smoothing pulls together the information on the mean function from neighbour locations. It is intuitively clear that there would not be any gains from such a ‘local pulling’ if the sample realisations of both $X_t(\cdot)$ and $\varepsilon_t(\cdot)$ are continuous in \mathbf{s} . The improvement may be obtained when the observations pulled together bring in different information, which is only possible when the sample realisations are discontinuous. The theory developed in next section indicates that the spatial smoothing may indeed improve the estimation when the ‘nugget effect’ is present, which results in the discontinuity in sample realisations as a function of \mathbf{s} . See also Lu *et al.* (2008) for more detailed discussions regarding the asymptotic theory with or without nugget effect.

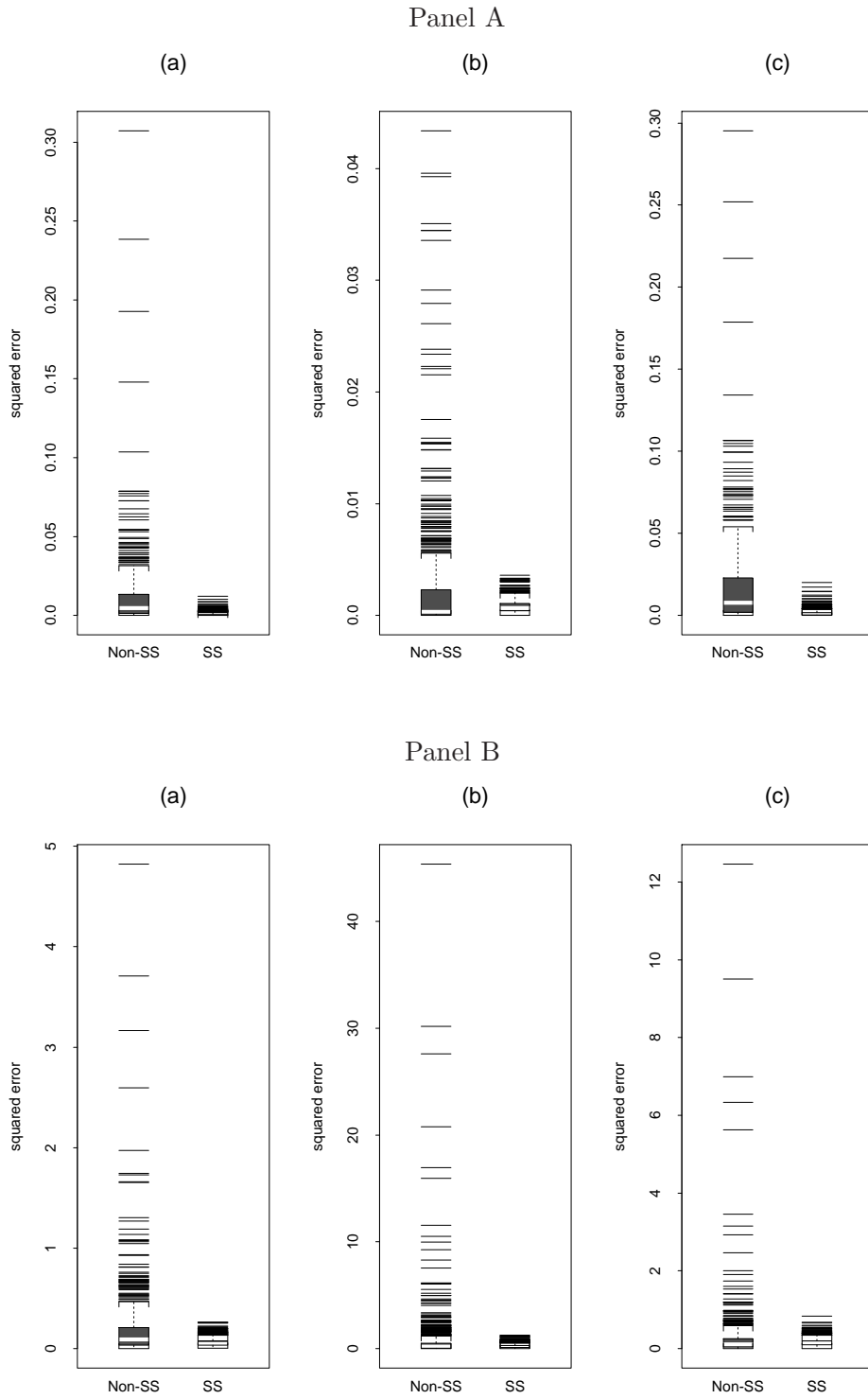


Figure 1: Simulation with $T = 60$ – Boxplots for the SEEs of the three components of α (Panel A), and the varying coefficient functions (Panel B): (a) $a(\cdot)$, (b) $b_1(\cdot)$, (c) $b_2(\cdot)$. The plots for the estimates defined in section 2.2.1 are labelled as ‘Non-SS’, and the plots for the spatial smoothing estimates are labelled as ‘SS’.

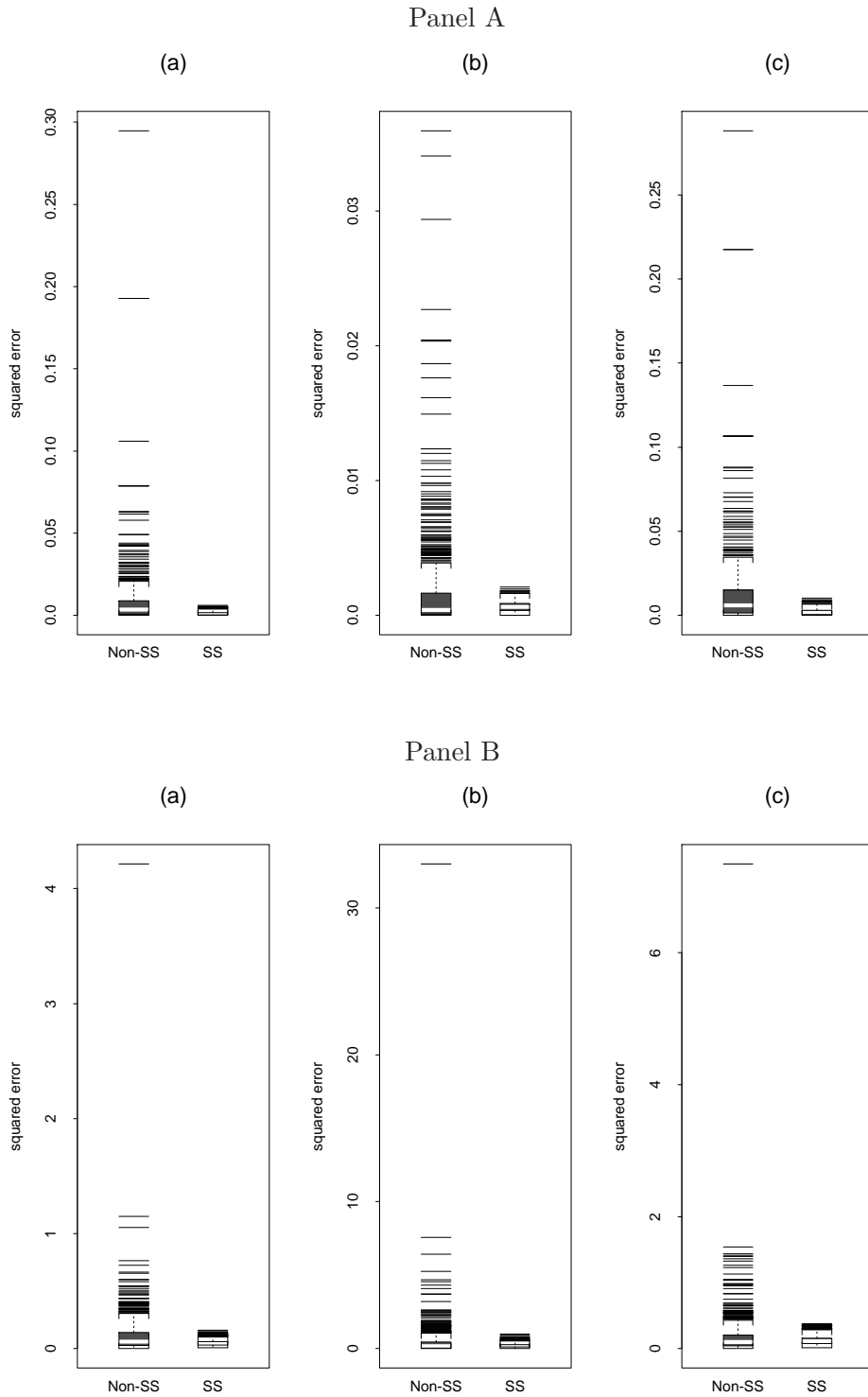


Figure 2: Simulation with $T = 100$ – Boxplots for the SEEs of the three components of α (Panel A), and the varying coefficient functions (Panel B): (a) $a(\cdot)$, (b) $b_1(\cdot)$, (c) $b_2(\cdot)$. The plots for the estimates defined in section 2.2.1 are labelled as ‘Non-SS’, and the plots for the spatial smoothing estimates are labelled as ‘SS’.

4 Asymptotic properties

The improvements due to the spatial smoothing have been illustrated via simulation in Section 3. In this section, we derive the asymptotic bias and variance of the estimators (2.9) – (2.11), which also show the benefits from spatial smoothing in the presence of the so-called nugget effect. Note that those asymptotic approximations are derived without the stationarity over space. On the other hand, the asymptotic normality for the time series estimators defined in Section 2.2.1 may be derived from Theorems 3.1 and 3.2 of Lu *et al.* (2007). We introduce some notation first.

We always assume the true index $\boldsymbol{\alpha}(\mathbf{s}) \in \mathbb{B}$ for all $\mathbf{s} \in \mathcal{S}$, where

$$\begin{aligned} \mathbb{B} = \{ & \boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^\tau \in \mathbb{R}^d : \|\boldsymbol{\beta}\| = 1, \text{ the first non-zero element is positive,} \\ & \text{and the last element is non-zero with } |\beta_d| \geq \epsilon_0 \}, \end{aligned} \quad (4.1)$$

and $\epsilon_0 > 0$ is a fixed constant. Put $\mathbb{X}_t(\mathbf{s}) = (1, \check{\mathbf{X}}_t(\mathbf{s})^\tau)^\tau = (1, X_{t1}(\mathbf{s}), \dots, X_{t,d-1}(\mathbf{s}))^\tau$, and

$$\mathbf{g}(\mathbf{s}, z, \boldsymbol{\beta}) = [E\{\mathbb{X}_t(\mathbf{s})\mathbb{X}_t^\tau(\mathbf{s})|\boldsymbol{\beta}^\tau \mathbf{X}_t(\mathbf{s}) = z\}]^{-1} E\{\mathbb{X}_t(\mathbf{s})Y_t(\mathbf{s})|\boldsymbol{\beta}^\tau \mathbf{X}_t(\mathbf{s}) = z\},$$

where the matrix inverse is ensured by Assumption C6 in the appendix. We denote the derivatives of \mathbf{g} as follows:

$$\dot{\mathbf{g}}_z(\mathbf{s}, z, \boldsymbol{\beta}) = \partial \mathbf{g}(\mathbf{s}, z, \boldsymbol{\beta}) / \partial z, \quad \dot{\mathbf{g}}_\beta(\mathbf{s}, z, \boldsymbol{\beta}) = \partial \mathbf{g}(\mathbf{s}, z, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}^\tau, \quad \ddot{\mathbf{g}}_z(\mathbf{s}, z, \boldsymbol{\beta}) = \partial^2 \mathbf{g}(\mathbf{s}, z, \boldsymbol{\beta}) / \partial z^2.$$

When $\boldsymbol{\beta} = \boldsymbol{\alpha}(\mathbf{s})$, we will often drop $\boldsymbol{\alpha}(\mathbf{s})$ in $\mathbf{g}\{\mathbf{s}, z, \boldsymbol{\alpha}(\mathbf{s})\}$ and its derivatives if no confusion arises. It therefore follows from (2.2) that $\mathbf{g}\{\mathbf{s}, z\}^\tau = \mathbf{g}\{\mathbf{s}, z, \boldsymbol{\alpha}(\mathbf{s})\}^\tau = \{a(\mathbf{s}, z), \mathbf{b}(\mathbf{s}, z)^\tau\}$. Further, we let $Z_t(\mathbf{s}) = \boldsymbol{\alpha}(\mathbf{s})^\tau \mathbf{X}_t(\mathbf{s})$, and define

$$\Gamma(\mathbf{s}) = E[\tilde{g}(\mathbf{s})^\tau \mathbb{X}_t(\mathbf{s}) \mathbb{X}_t(\mathbf{s})^\tau \tilde{g}(\mathbf{s}) w\{Z_t(\mathbf{s})\}], \quad (4.2)$$

where $\tilde{g}(\mathbf{s}) = \dot{\mathbf{g}}_z\{\mathbf{s}, Z_t(\mathbf{s}), \boldsymbol{\alpha}(\mathbf{s})\} \mathbf{X}_t(\mathbf{s})^\tau + \dot{\mathbf{g}}_\beta\{\mathbf{s}, Z_t(\mathbf{s}), \boldsymbol{\alpha}(\mathbf{s})\}$.

Let $\alpha_j(\mathbf{s})$ and $\ddot{\alpha}_j(\mathbf{s})$ be the j -th component of $\boldsymbol{\alpha}(\mathbf{s})$ and its second order derivative matrix with respect to \mathbf{s} , respectively. Recall that $h = h_T$ and $\bar{h} = \bar{h}_T$ are the two bandwidths used for time series smoothing in Section 2.2.1, and $\tilde{h} = \tilde{h}_N$ is the bandwidth for spatial smoothing in Section 2.2.2. Put $\mu_{2,K} = \int u^2 K(u) du$, $\boldsymbol{\mu}_{2,W} = \int_{\mathcal{S}} \mathbf{u} \mathbf{u}^\tau W(\mathbf{u}) d\mathbf{u}$ and $\nu_{0,K} = \int K^2(u) du$, where K and W are the two kernel functions used in our estimation.

Now we are ready to present the asymptotic biases and variances for the spatial smoothing estimators $\tilde{\boldsymbol{\alpha}}(\mathbf{s}_0)$, $\tilde{a}(\mathbf{s}_0)$ and $\tilde{b}(\mathbf{s}_0)$, which are derived when $T \rightarrow \infty$. The key is to show that the

time series estimators $\widehat{\boldsymbol{\alpha}}(\mathbf{s})$, $\widehat{a}(\mathbf{s})$ and $\widehat{b}(\mathbf{s})$ converge uniformly in \mathbf{s} over a small neighbourhood of \mathbf{s}_0 , which is established using the results in Lu *et al.* (2007). Naturally the derived asymptotic approximations for the biases and variances of $\widetilde{\boldsymbol{\alpha}}(\mathbf{s}_0)$, $\widetilde{a}(\mathbf{s}_0)$ and $\widetilde{b}(\mathbf{s}_0)$ depend on N , and those approximations admit more explicit expressions when $N \rightarrow \infty$. Note we write $a_N \simeq b_N$ if $\lim_{N \rightarrow \infty} a_N/b_N = 1$.

Theorem 4.1 *Let conditions C1 – C10 listed in the Appendix hold. Assume that $\mathbf{s}_0 \in \mathcal{S}$ and $f(\mathbf{s}_0) > 0$, where $f(\cdot)$ is given in condition C8. Then as $T \rightarrow \infty$, it holds that*

$$\begin{aligned} \widetilde{\boldsymbol{\alpha}}(\mathbf{s}_0) - \boldsymbol{\alpha}(\mathbf{s}_0) &= \mathcal{B}_{1,N}(\mathbf{s}_0)h^2(1 + o_P(1)) + \mathcal{B}_{2,N}(\mathbf{s}_0) \\ &+ T^{-1/2} \{ \mathcal{V}_{1,N}(\mathbf{s}_0) + \mathcal{V}_{2,N}(\mathbf{s}_0) \} \boldsymbol{\eta}(\mathbf{s}_0)(1 + o_P(1)), \end{aligned}$$

where $\boldsymbol{\eta}(\mathbf{s}_0)$ is a $d \times 1$ random vector with zero mean and identity covariance matrix,

$$\mathcal{B}_{1,N}(\mathbf{s}) \simeq -\frac{1}{2}\Gamma^{-1}(\mathbf{s})E[\widetilde{g}(\mathbf{s})^\tau \mathbb{X}_t(\mathbf{s})\mathbb{X}_t(\mathbf{s})^\tau \ddot{\mathbf{g}}_z\{\mathbf{s}, Z_t(\mathbf{s}), \boldsymbol{\alpha}(\mathbf{s})\}w\{Z_t(\mathbf{s})\}]\mu_{2,K}, \quad (4.3)$$

$$\mathcal{B}_{2,N}(\mathbf{s}) \simeq \frac{1}{2}\widetilde{h}^2 (\text{tr} [\ddot{\boldsymbol{\alpha}}_1(\mathbf{s})\mu_{2,W}], \dots, \text{tr} [\ddot{\boldsymbol{\alpha}}_d(\mathbf{s})\mu_{2,W}])^\tau, \quad (4.4)$$

and $\mathcal{V}_{1,N}(\mathbf{s}_0)$ and $\mathcal{V}_{2,N}(\mathbf{s}_0)$ are two $d \times d$ matrices, satisfying

$$\mathcal{V}_{1,N}(\mathbf{s})\mathcal{V}_{1,N}^\tau(\mathbf{s}) \simeq (\sigma_1^2(\mathbf{s}) + \sigma_2^2(\mathbf{s}))\Gamma^{-1}(\mathbf{s})\mathbf{V}(\mathbf{s}, \mathbf{s})\Gamma^{-1}(\mathbf{s}) \left\{ \frac{1}{N\widetilde{h}^2} f(\mathbf{s})^{-1} \int W^2(\mathbf{u})d\mathbf{u} \right\}, \quad (4.5)$$

$$\mathcal{V}_{2,N}(\mathbf{s})\mathcal{V}_{2,N}^\tau(\mathbf{s}) \simeq \sigma_1^2(\mathbf{s})\Gamma^{-1}(\mathbf{s})\mathbf{V}_1(\mathbf{s}, \mathbf{s})\Gamma^{-1}(\mathbf{s}). \quad (4.6)$$

In the above expressions, $\Gamma(\mathbf{s})$ is defined in (4.2), and $\sigma_1^2(\mathbf{s})$ and $\sigma_2^2(\mathbf{s})$ are defined in C2, and \mathbf{V} and \mathbf{V}_1 in C3 in the Appendix.

Let $\widetilde{\vartheta}(\mathbf{s}_0, z) \equiv (\widetilde{a}(\mathbf{s}_0, z), \widetilde{\mathbf{b}}(\mathbf{s}_0, z)^\tau)^\tau$ and $\vartheta(\mathbf{s}, z) \equiv (a(\mathbf{s}, z), \mathbf{b}(\mathbf{s}, z)^\tau)^\tau$. Denote by $\vartheta_j(\mathbf{s}, z)$ the j -th component of $\vartheta(\mathbf{s}, z)$, and $\ddot{\vartheta}_{\mathbf{s},j}(\mathbf{s}, z)$ its second derivative matrix with respect to \mathbf{s} . Denote by $\ddot{\vartheta}_z(\mathbf{s}, z) = (\ddot{a}_z(\mathbf{s}, z), \ddot{\mathbf{b}}_z(\mathbf{s}, z)^\tau)^\tau$ the second order derivatives with respect to z .

Theorem 4.2 *Let the conditions of Theorem 4.1 hold, and $\bar{h} = O(T^{-1/5})$. Let the density function of $\boldsymbol{\alpha}(\mathbf{s}_0)^\tau \mathbf{X}_t(\mathbf{s}_0)$ be positive at z . Then as $T \rightarrow \infty$, it holds that*

$$\begin{aligned} \widetilde{\vartheta}(\mathbf{s}_0, z) - \vartheta(\mathbf{s}_0, z) &= \mathcal{B}_{3,N}(\mathbf{s}_0, z)\bar{h}^2(1 + o_P(1)) + \mathcal{B}_{4,N}(\mathbf{s}_0, z) \\ &+ \left\{ (T\bar{h})^{-1/2}\mathcal{V}_{3,N} + T^{-1/2}\mathcal{V}_{4,N} \right\} \boldsymbol{\xi}(\mathbf{s}_0)(1 + o_P(1)), \end{aligned}$$

where $\xi(\mathbf{s}_0)$ is a $d \times 1$ random vector with zero mean and identity covariance matrix,

$$\mathcal{B}_{3,N}(\mathbf{s}_0, z) \simeq \frac{1}{2} \mu_{2K} \ddot{\vartheta}_z(\mathbf{s}_0, z), \quad (4.7)$$

$$\mathcal{B}_{4,N}(\mathbf{s}_0, z) \simeq \frac{1}{2} \tilde{h}^2 \left(\text{tr} \left[\ddot{\vartheta}_{\mathbf{s},1}(\mathbf{s}_0, z) \mu_{2,W} \right], \dots, \text{tr} \left[\ddot{\vartheta}_{\mathbf{s},d}(\mathbf{s}_0, z) \mu_{2,W} \right] \right)^\tau, \quad (4.8)$$

and $\mathcal{V}_{3,N}$ and $\mathcal{V}_{4,N}$ are two $d \times d$ matrices, satisfying

$$\mathcal{V}_{3,N} \mathcal{V}_{3,N}^\tau \simeq (\sigma_1^2(\mathbf{s}_0) + \sigma_2^2(\mathbf{s}_0)) \nu_{0,K} \mathbf{A}^{-1}(\mathbf{s}_0, z) f_Z^{-1}(\mathbf{s}_0, z) \left\{ \frac{1}{N \tilde{h}^2} f^{-1}(\mathbf{s}_0) \int W^2(\mathbf{s}) d\mathbf{s} \right\}, \quad (4.9)$$

$$\mathcal{V}_{4,N} \mathcal{V}_{4,N}^\tau \simeq \sigma_1^2(\mathbf{s}_0) (\mathbf{A}^{-1}(\mathbf{s}_0, z) \mathbf{A}_1(\mathbf{s}_0, \mathbf{s}_0; z, z) \mathbf{A}^{-1}(\mathbf{s}_0, z)) f_Z^{-2}(\mathbf{s}_0, z) f^*(\mathbf{s}_0, z). \quad (4.10)$$

where $\sigma_1^2(\mathbf{s})$ and $\sigma_2^2(\mathbf{s})$ are defined in (C2), and $\mathbf{A}(\mathbf{s}, z) = \mathbf{A}(\mathbf{s}, \mathbf{s}, z, z)$ and $\mathbf{A}_1(\mathbf{s}, \mathbf{s}, z, z)$ in (C5) in the Appendix.

Remark. (i) In the above theorems, (4.3), (4.4), (4.7) and (4.8) are approximate biases whereas (4.5), (4.6), (4.9) and (4.10) are approximate variances.

(ii) In the presence of nugget effect specified in C2 and C3 in the Appendix, the spatial smoothing estimators $\tilde{\alpha}(\mathbf{s}_0)$, $\tilde{a}(\mathbf{s}_0, z)$ and $\tilde{\mathbf{b}}(\mathbf{s}_0, z)$ have smaller asymptotic variances than the corresponding time series estimators $\hat{\alpha}(\mathbf{s}_0)$, $\hat{a}(\mathbf{s}_0, z)$ and $\hat{\mathbf{b}}(\mathbf{s}_0, z)$. In fact using the results in Lu *et al.* (2007), we may deduce that

$$\hat{\alpha}(\mathbf{s}) - \alpha(\mathbf{s}) = \mathcal{B}_1(\mathbf{s}) h^2 (1 + o_P(1)) + T^{-1/2} \mathcal{A}_1(\mathbf{s}) \eta(\mathbf{s}) (1 + o_P(1)), \quad (4.11)$$

where $\mathcal{B}_1(\mathbf{s})$ is as defined in Theorem 4.1, and

$$\mathcal{A}_1(\mathbf{s}) \mathcal{A}_1^\tau(\mathbf{s}) = (\sigma_1^2(\mathbf{s}) + \sigma_2^2(\mathbf{s})) \Gamma^{-1}(\mathbf{s}) \mathbf{V}(\mathbf{s}, \mathbf{s}) \Gamma^{-1}(\mathbf{s}),$$

and that

$$\hat{\vartheta}(\mathbf{s}, z) - \vartheta(\mathbf{s}, z) = \mathcal{B}_3(\mathbf{s}, z) \bar{h}^2 (1 + o_P(1)) + (T \bar{h})^{-1/2} \mathcal{A}_3(\mathbf{s}, z) \xi(\mathbf{s}) (1 + o_P(1)), \quad (4.12)$$

where $\mathcal{B}_3(\mathbf{s}, z)$ is as defined in Theorem 4.2, and

$$\mathcal{A}_3(\mathbf{s}, z) \mathcal{A}_3^\tau(\mathbf{s}, z) = (\sigma_1^2(\mathbf{s}) + \sigma_2^2(\mathbf{s})) \nu_{0,K} \mathbf{A}^{-1}(\mathbf{s}, z) \{f_Z(\mathbf{s}, z)\}^{-1}.$$

Comparing (4.11) and (4.12) with Theorems 4.1 and 4.2, respectively, we note that both (4.5) and

(4.9) tend to 0, and the asymptotic variances of the spatial smoothing estimators are therefore much smaller.

(iii) In the case of no nugget effect (*i.e.* $\sigma_2^2(\mathbf{s}_0) = 0$ in C2, $\mathbf{A}_0(\mathbf{s}_0, z) = \mathbf{A}_2(\mathbf{s}_0, \mathbf{s}_0, z, z) \equiv 0$ and $\mathbf{V}_2(\mathbf{s}_0, \mathbf{s}_0) = 0$ in C3), spatial smoothing cannot reduce the asymptotic variance of the time series estimators $\hat{\boldsymbol{\alpha}}(\mathbf{s}_0)$, $\hat{\boldsymbol{\alpha}}(\mathbf{s}_0)$ and $\hat{\mathbf{b}}(\mathbf{s}_0)$. See also Lu *et al.* (2008). This is due to the fact that the spatial smoothing uses effectively the data at locations within a distance \tilde{h} from \mathbf{s}_0 . Due to the continuity of the function $\gamma_1(\cdot, \cdot)$ stated in C2, all the $\varepsilon_t(\mathbf{s})$'s from those locations are asymptotically identical. We argue that asymptotic theory under this setting presents an excessively gloomy picture. Adding a nugget effect in the model brings the theory closer to reality since in practice the data used in local spatial smoothing usually contain some noise components which are not identical even within a very small neighborhood. Note that the nugget effect is not detectable in practice since we can never estimate $\gamma(\mathbf{s} + \Delta, \mathbf{s})$ defined in (6.1) below for $\|\Delta\|$ less than the minimum pairwise-distance among observed locations. See also Remark 3 of Zhang *et al.* (2003).

The proofs of Theorems 4.1 and 4.2 are included in an extended version of this paper available at <http://stats.lse.ac.uk/q.yao/qyao.links/paper/spatioVLM.pdf>.

5 A real data example

We illustrate the proposed methodology with a meteorological data set.

5.1 Data

We analyze the daily mean sea level pressure (MSLP) readings measured in units of Pascal in an area of the North Sea with longitudes from 20 degrees East to 20 degrees West and latitudes from 50 to 60 degrees North. This area is heavily influenced by travelling low pressure systems. The grid points are of size 2.5×2.5 degrees with total number of $17 \times 5 = 85$ spatial locations. The time period is winter 2001-2002 with 100 daily observations at each location starting from 1 December 2001. The data are provided by the National Centers for Environmental Prediction-National Center for Atmospheric Research (NCEP-NCAR). This type of data is commonly used in climate analysis, and more detailed information about the data can be found in Kalnay *et al.* (1996). Trends are not removed, and therefore the original time series of mean sea level pressure (MSLP) at each location is generally non-stationary, and we have chosen to work with the differenced

series (daily change series). Figure 3 displays some plots for MSLP data at five randomly selected locations. We denote the daily changes of MSLP as $Y_t(\mathbf{s}_{ij}) = Y_t(u_i, v_j)$, $t = 1, \dots, 99$, $u_i = 60 - (i - 1) \times 2.5$ with $i = 1, \dots, 5$, and $v_j = -17.5 + (j - 1) \times 2.5$ with $j = 1, 2, \dots, 17$. From the middle column of Figure 3, the daily change series of MSLP, $Y_t(\mathbf{s})$, looks approximately stationary at each site.

The contour plots of the daily changes are given in Figure 4 for the time period $t = 1$ to $t = 20$. These plots show the difference from one day to another as a function of spatial coordinate. For the high pressure period (approximately $t = 8$ to $t = 16$) it is seen that for most spatial locations there are small changes, corresponding to the fact that a high pressure system is fairly stable in time and in space. The zero-contour of no change is also seen to be fairly stable running in the north-south direction roughly in the middle of this area. Just prior to $t = 8$, positive pressure gradients are dominating corresponding to the build-up of the high pressure system, whereas negative gradients are dominating after $t = 16$, when the high pressure system is deteriorating.

5.2 Varying coefficient modelling

Now we model the daily changes $Y_t(\mathbf{s})$ by the varying coefficient model (1.1) with

$$\mathbf{X}_t(\mathbf{s}_{ij})^\tau = (Y_{t-1}(\mathbf{s}_{i-1,j}), Y_{t-1}(\mathbf{s}_{i+1,j}), Y_{t-1}(\mathbf{s}_{i,j-1}), Y_{t-1}(\mathbf{s}_{i,j+1}), Y_{t-1}(\mathbf{s}_{ij})), \quad (5.1)$$

i.e. $\mathbf{X}(\mathbf{s}_{ij})$ consists of the lagged values at the four nearest neighbours of the location \mathbf{s}_{ij} , and the lagged value at \mathbf{s}_{ij} itself. We only use the time lag 1 since the autocorrelation of the daily change data is weak. It is clear that this specification does not require that the data are recorded regularly over the space. With $\mathbf{X}(\mathbf{s}_{ij})$ specified above, (1.1) is now of the form

$$Y_t(\mathbf{s}_{ij}) = a(\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})) + b_1(\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij}))Y_{t-1}(\mathbf{s}_{i-1,j}) + b_2(\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij}))Y_{t-1}(\mathbf{s}_{i+1,j}) \\ + b_3(\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij}))Y_{t-1}(\mathbf{s}_{i,j-1}) + b_4(\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij}))Y_{t-1}(\mathbf{s}_{i,j+1}) + b_5(\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij}))Y_{t-1}(\mathbf{s}_{ij}) + \varepsilon_t(\mathbf{s}_{ij}) \quad (5.2)$$

with $Z_t(\mathbf{s}_{ij}) = \boldsymbol{\alpha}(\mathbf{s}_{ij})^\tau \mathbf{X}_t(\mathbf{s}_{ij})$, $\boldsymbol{\alpha}(\mathbf{s}_{ij})^\tau = (\alpha_1(\mathbf{s}_{ij}), \alpha_2(\mathbf{s}_{ij}), \alpha_3(\mathbf{s}_{ij}), \alpha_4(\mathbf{s}_{ij}), \alpha_5(\mathbf{s}_{ij}))$ and $\mathbf{b}^\tau = (b_1, b_2, b_3, b_4, b_5)$. The estimation was based on data from 45 locations: $Y_t(\mathbf{s}_{ij}) = Y_t(u_i, v_j)$, $t = 1, \dots, 99$, $i = 2, 3, 4$, and $j = 2, \dots, 16$, owing to the boundary effect in space. The bandwidths were selected using the methods specified in Section 2.3, in which we let $q = 10$ and take $\bar{h}_k = C_T \bar{h}_{ck}$ with $C_T = \text{std}(Z_t)(99)^{-1/5}$ and $\text{std}(Z_t)$ being the sample standard deviation of $\{Z_t = \hat{\boldsymbol{\alpha}}^\tau \mathbf{X}_t(\mathbf{s})\}_{t=1}^{99}$ at location \mathbf{s} , and $\bar{h}_{ck} = 0.1k$ for $k = 1, 2, \dots, q$. The estimated $a(z)$, $b_1(z)$, $b_2(z)$, $b_3(z)$, $b_4(z)$ and $b_5(z)$ at 45 locations, together with their averages (over the 45 locations), are plotted in Figure 5 without spatial smoothing, and Figure 6 with the spatial smoothing.

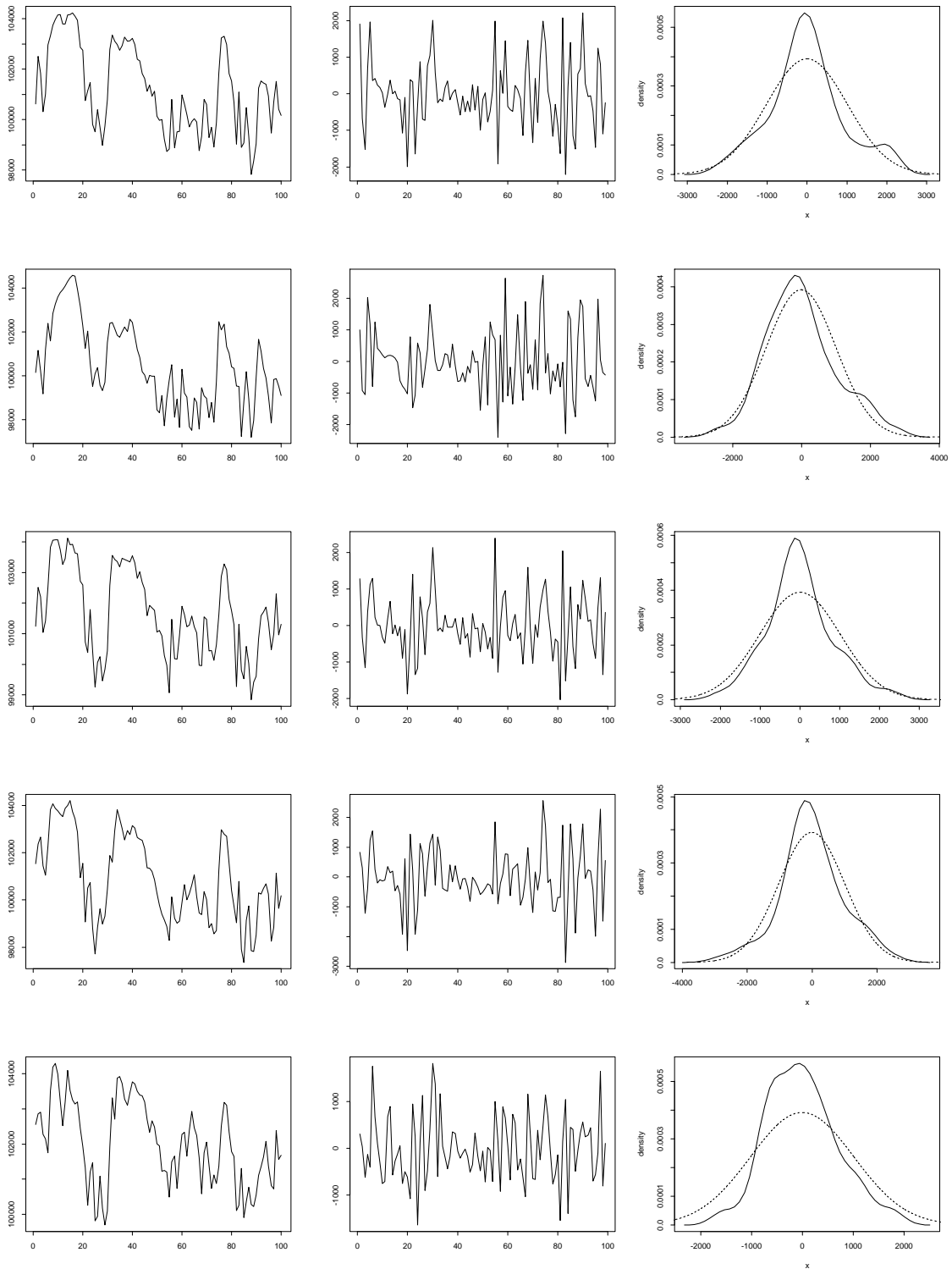


Figure 3: Modelling of MSLP data – Column on the left: the time series plots of the daily MSLP series in 1 December 2001 – 10 March 2002. Middle column: the daily changes of MSLP. Column on the right: the estimated density function (solid curves) of the daily changes together with the normal density with the same mean and variance (dashed curves).

If the meteorological data could be described linearly, the functions displayed in Figures 5 and 6 would be constant. But clearly they are not. If these plots are continued for higher and lower values of z , the asymptote turns out to be roughly horizontal. This means that in a low pressure situation the system behaves roughly as a linear system, but not in the high pressure zones ($z \approx 0$) and the transition zones. This indicates that altogether the weather system is nonlinear. In another context it would be of interest to try to give a detailed meteorological interpretation of the curves. But since this is a general paper, we satisfy ourselves by noting the following: 1) $z \approx 0$: This corresponds to a high pressure situation primarily. The curves for b_i , $i = 1, \dots, 5$, have their maximum/minimum at zero, the negative and positive peak values roughly balancing out, and given that the \mathbf{X}_t -variables stay around zero for a high pressure situation, the effect is that Y_t stays around zero, and the high pressure continues. 2) $z < 0$: This is the situation when a high pressure deteriorates, or when new low pressure areas are coming in. It is seen that there are only small contributions from the north (b_1), east (b_4) and the site itself (b_5), but a main positive contribution from west (b_3) having the effect of maintaining a negative difference in X_t . It is tempting to interpret this as having to do with low pressure systems coming primarily from the west in the North Sea.

The contour plots (not shown) of the estimated index $\boldsymbol{\alpha}^T \mathbf{X}(\mathbf{s})$ resemble the patterns of Figure 4; indicating that it catches the major spatial variation of $Y(\mathbf{s})$. Furthermore there were little variation in the index vector $\boldsymbol{\alpha}(\mathbf{s})$ over the space. The average value is $\bar{\boldsymbol{\alpha}}^T = (-0.367, -0.252, 0.174, 0.219, 0.851)$. Note that the last component of the index vector is substantially greater than all the others, indicating the dominating role of the lagged value from the same site; see (5.1). We plot the 45 estimated index time series $Z_t(\mathbf{s})$ in Figure 7(a), together with the first principal component series $\gamma(t) = \boldsymbol{\lambda}_1^T \mathbf{Y}_t$ where $\boldsymbol{\lambda}_1$ is the first principal component of the data (explaining about 46% of the total variation). Note that the oscillatory patterns of $\gamma(t)$ and the estimated index series $Z_t(\mathbf{s})$ are similar. The high pressure areas can again be identified as regions where the $Z_t(\mathbf{s})$ -series are close to zero, building up a high pressure area corresponds to positive $Z_t(\mathbf{s})$ (but not conversely), and deterioration to negative $Z_t(\mathbf{s})$ (but not conversely). There does not seem to be a corresponding regularity for the $\gamma(t)$ series, which is sometimes taken as a representative index for the North Atlantic Oscillation (NAO) on another time scale.

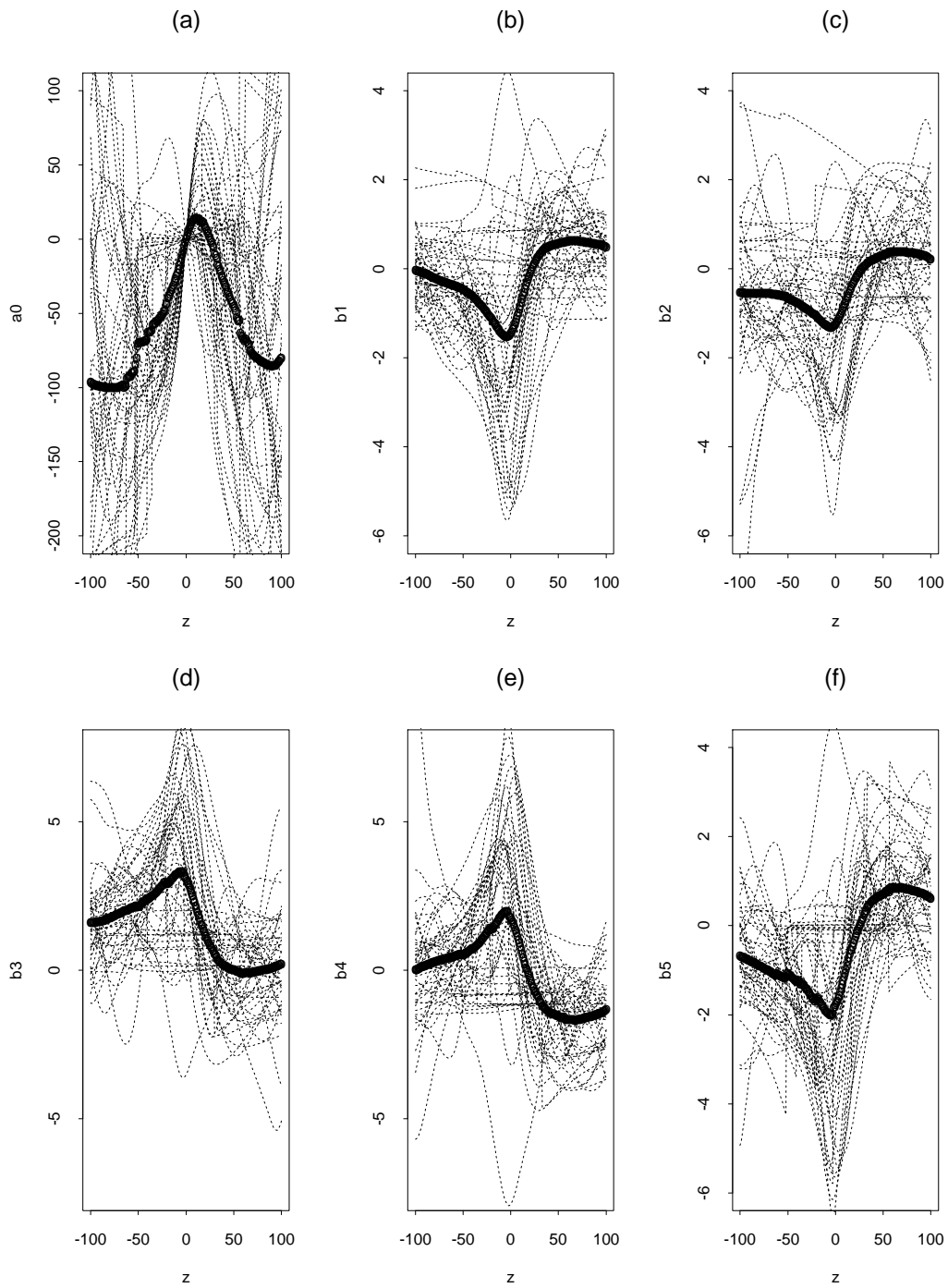


Figure 5: Modelling of MSLP data – The estimated curves (dotted lines) of 45 locations together with the averaged curve (solid line) for (a) $a(\cdot)$, (b) $b_1(\cdot)$, (c) $b_2(\cdot)$, and (d) $b_3(\cdot)$, (e) $b_4(\cdot)$, and (f) $b_5(\cdot)$. No spatial smoothing applied in estimation.

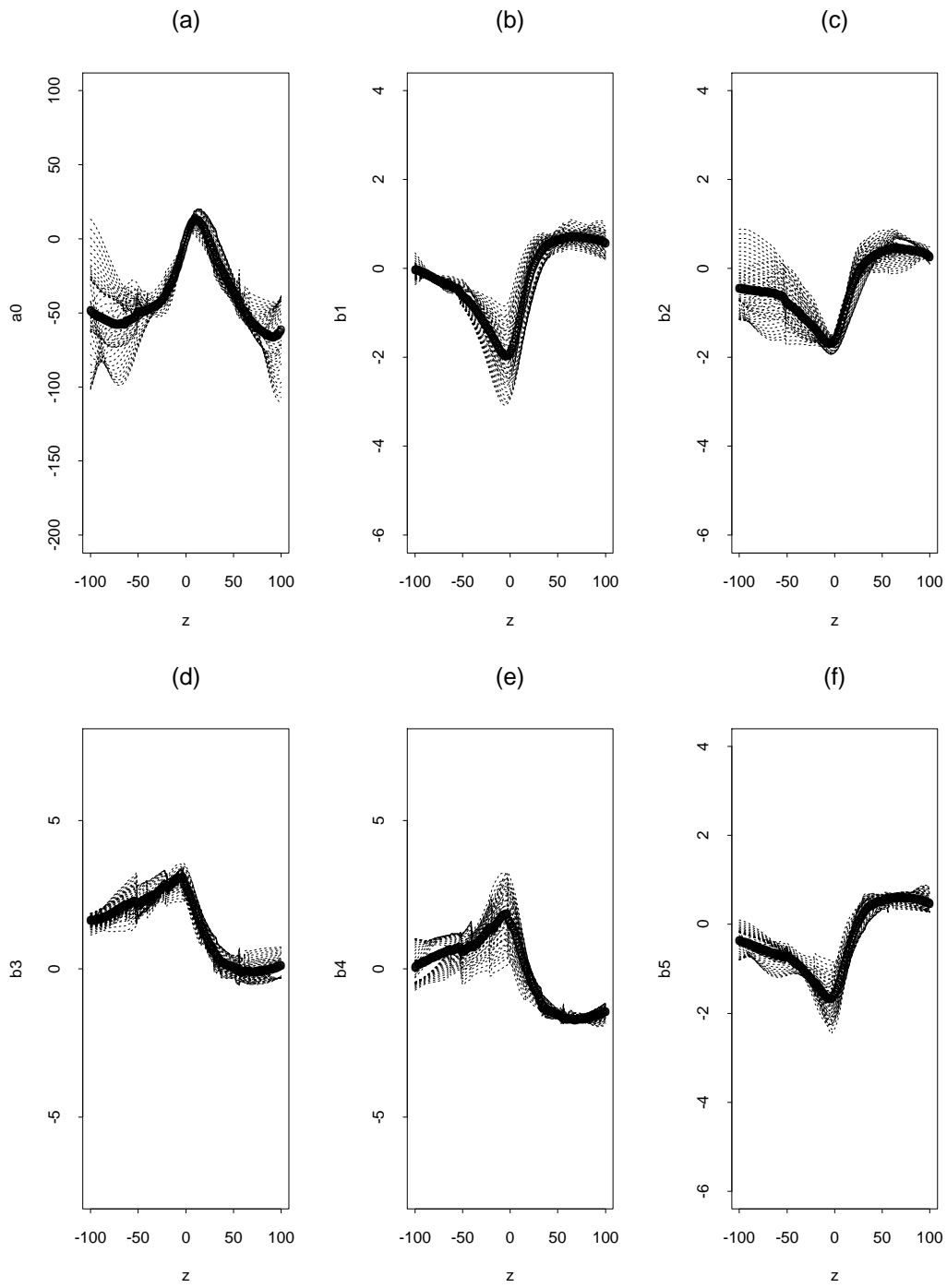


Figure 6: Modelling of MSLP data – The spatial-smoothed estimated curves (dotted lines) of 45 locations together with the averaged curve (solid line) for (a) $a(\cdot)$, (b) $b_1(\cdot)$, (c) $b_2(\cdot)$, and (d) $b_3(\cdot)$, (e) $b_4(\cdot)$, and (f) $b_5(\cdot)$.

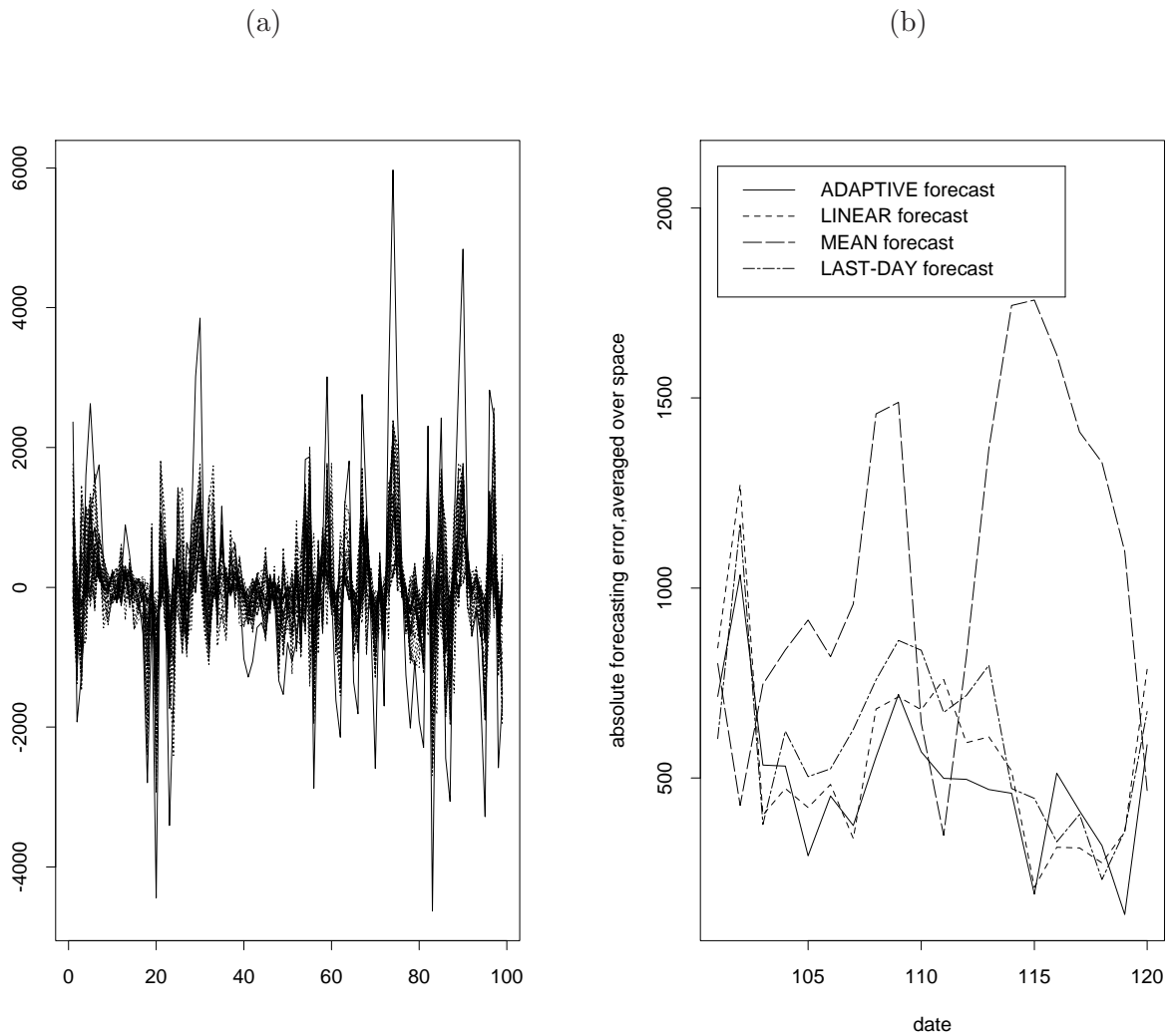


Figure 7: *MSLP data – (a) Time series plots of the 45 estimated index series (dotted lines) and the global spatial index series (i.e. the first principle component series, solid line, scaled down by 3). (b) Absolute one-step ahead forecasting errors, averaged over 45 locations, over the period of 11 – 30 March 2002.*

5.3 Forecasting comparisons

To further assess the capability of model (5.2), we compared its post-sample forecasting performance with those of other models including linear one. We used the adaptive varying-coefficient model (AVCM) fitted from the above for one-step ahead prediction for the 20 MSLP values in the period of 11 – 30 March 2002 over the 45 locations. Also included in the comparison are the LINEAR model which may be viewed as a special case of AVCM with constant coefficient

Table 1: Mean absolute predictive errors (MAPE) of the four different models

Model	AVCM	LINEAR	MEAN	LAST-DAY
MAPE	494.1	553.0	1052.8	599.9

Table 2: Mean absolute predictive errors (MAPE) of AVCM with different bandwidths

(a) $\bar{h}_c = \bar{h}_{0c}$, $\tilde{h} = \tilde{h}_0$:

h	$h_0 - 4$	$h_0 - 1$	$h_0 + 1$	$h_0 + 4$
MAPE	499.9	496.8	498.9	508.7

(b) $h = h_0$, $\tilde{h} = \tilde{h}_0$:

\bar{h}_c	$\bar{h}_{0c} - 0.4$	$\bar{h}_{0c} - 0.1$	$\bar{h}_{0c} + 0.1$	$\bar{h}_{0c} + 0.4$
MAPE	523.3	494.1	496.2	495.0

(c) $h = h_0$, $\bar{h}_c = \bar{h}_{0c}$:

\tilde{h}	$\tilde{h}_0 - 0.4$	$\tilde{h}_0 - 0.1$	$\tilde{h}_0 + 0.1$	$\tilde{h}_0 + 0.4$
MAPE	505.2	495.3	493.9	493.3

functions, the MEAN forecast which forecasts the future based on the mean of the past values from 1 December 2001 onwards, and the LAST-DAY forecast which forecasts tomorrow's value by today's value. The absolute forecasting errors, averaged over the 45 space locations, are plotted in Figure 7(b). The mean absolute predictive errors (MAPE), over the 45 space locations and the 20 days of 11 – 30 March 2002, are listed in Table 1. Overall our adaptive varying-coefficient model outperforms the other three simple models. This lends further support to the assertion that the dynamic structure of the MSLP is nonlinear and it cannot be accommodated, for example, in a linear dynamic model.

In the above comparison, the bandwidths used in fitting AVCM were selected using the methods specified in Section 2.3. Over the 45 locations, the selected values for h , \bar{h}_c in $\bar{h} = C_T \bar{h}_c$ and \tilde{h} are around, respectively, 15, 0.6 and 0.7. To examine the sensitivity of the AVCM forecasting on the bandwidths, we repeat the above exercises with the bandwidths shifted from the selected values. We denote by h_0 , \bar{h}_{0c} and \tilde{h}_0 the selected values of h , \bar{h}_c and \tilde{h} using the methods specified in Section 2.3. Table 2 lists the MAPE of the AVCM with different bandwidths. Comparing with Table 1, the MAPE of the AVCM varies with the bandwidths used in the estimation, but the

variation is not excessive. Further the AVCM always offers better prediction than the three other models.

6 Appendix: Regularity conditions

We list below the regularity conditions, among which C4(i,ii), C5-C7 and C9-C10 are basically inherited from Lu, Tjøstheim and Yao (2007) in the time series context.

- C1. For each \mathbf{s} , $\{\varepsilon_t(\mathbf{s}), t \geq 1\}$ is a sequence of independent and identically distributed random variables. The distribution of $\varepsilon_t(\mathbf{s})$ is the same for all $\mathbf{s} \in \mathcal{S}$. Further, for each $t > 1$, $\{\varepsilon_t(\mathbf{s}), \mathbf{s} \in \mathcal{S}\}$ is independent of $\{(Y_{t-j}(\mathbf{s}), \mathbf{X}_{t+1-j}(\mathbf{s})), \mathbf{s} \in \mathcal{S} \text{ and } j \geq 1\}$. The spatial covariance function

$$\gamma(\mathbf{s}_1, \mathbf{s}_2) \equiv \text{Cov}\{\varepsilon_t(\mathbf{s}_1), \varepsilon_t(\mathbf{s}_2)\} \quad (6.1)$$

is bounded over \mathcal{S}^2 .

- C2. For any $t \geq 1$ and $\mathbf{s} \in \mathcal{S}$,

$$\varepsilon_t(\mathbf{s}) = \varepsilon_{1,t}(\mathbf{s}) + \varepsilon_{2,t}(\mathbf{s}) \quad (6.2)$$

where $\{\varepsilon_{1,t}(\mathbf{s}), t \geq 1, \mathbf{s} \in \mathcal{S}\}$ and $\{\varepsilon_{2,t}(\mathbf{s}), t \geq 1, \mathbf{s} \in \mathcal{S}\}$ are two independent processes, and both fulfill the conditions imposed on $\varepsilon_t(\mathbf{s})$ in C1 above. Further, $\gamma_1(\mathbf{s}_1, \mathbf{s}_2) \equiv \text{Cov}\{\varepsilon_{1,t}(\mathbf{s}_1), \varepsilon_{1,t}(\mathbf{s}_2)\}$ is continuous in $(\mathbf{s}_1, \mathbf{s}_2)$ (we will denote $\sigma_1^2(\mathbf{s}_1) = \gamma_1(\mathbf{s}_1, \mathbf{s}_1)$), and $\gamma_2(\mathbf{s}_1, \mathbf{s}_2) \equiv \text{Cov}\{\varepsilon_{2,t}(\mathbf{s}_1), \varepsilon_{2,t}(\mathbf{s}_2)\} = 0$ if $\mathbf{s}_1 \neq \mathbf{s}_2$, and $\sigma_2^2(\mathbf{s}) = \gamma_2(\mathbf{s}, \mathbf{s}) > 0$ is continuous.

- C3. (i) $\mathbf{A}(\mathbf{s}, \mathbf{s}'; z_1, z_2) \equiv E\{\mathbb{X}_t(\mathbf{s})\mathbb{X}_t(\mathbf{s}')^\tau | Z_t(\mathbf{s}) = z_1, Z_t(\mathbf{s}') = z_2\} = \mathbf{A}_1(\mathbf{s}, \mathbf{s}'; z_1, z_2) + \mathbf{A}_2(\mathbf{s}, \mathbf{s}'; z_1, z_2)$, where $\mathbf{A}_1(\mathbf{s}, \mathbf{s}'; z_1, z_2)$ is continuous, and $\mathbf{A}_2(\mathbf{s}, \mathbf{s}; z_1, z_2)$ is a positive definite matrix. Further, $\mathbf{A}_2(\mathbf{s}, \mathbf{s}'; z_1, z_2) = 0$ if $\mathbf{s} \neq \mathbf{s}'$, and $\mathbf{A}_0(\mathbf{s}, z) \equiv \mathbf{A}_2(\mathbf{s}, \mathbf{s}; z, z)$ is continuous.

(ii) Set $\mathbb{V}_t(\mathbf{s}) = \mathbb{W}_t(\mathbf{s}) - \mathbf{B}_t(\mathbf{s})\mathbb{X}_t(\mathbf{s})$, where $\mathbf{B}_t(\mathbf{s}) = E\{\mathbb{W}_t(\mathbf{s})\mathbb{X}_t(\mathbf{s})^\tau | Z_t(\mathbf{s})\} [E\{\mathbb{X}_t(\mathbf{s})\mathbb{X}_t(\mathbf{s})^\tau | Z_t(\mathbf{s})\}]^{-1}$ and $\mathbb{W}_t(\mathbf{s}) = \mathbf{X}_t(\mathbf{s})\dot{\mathbf{g}}_z(\{\mathbf{s}, Z_t(\mathbf{s})\}^\tau \mathbb{X}_t(\mathbf{s}))$. Then $\mathbf{V}(\mathbf{s}_1, \mathbf{s}_2) \equiv E[\mathbb{V}_t(\mathbf{s}_1)\mathbb{V}_t(\mathbf{s}_2)^\tau w\{Z_t(\mathbf{s}_1)\}w\{Z_t(\mathbf{s}_2)\}] = \mathbf{V}_1(\mathbf{s}_1, \mathbf{s}_2) + \mathbf{V}_2(\mathbf{s}_1, \mathbf{s}_2)$, where $\mathbf{V}_1(\mathbf{s}_1, \mathbf{s}_2)$ is continuous, $\mathbf{V}_2(\mathbf{s}_1, \mathbf{s}_2) = 0$ if $\mathbf{s}_1 \neq \mathbf{s}_2$, and $\mathbf{V}_2(\mathbf{s}, \mathbf{s})$ is positive-definite and continuous.

- C4. (i) $\mathbf{g}(\mathbf{s}, z, \boldsymbol{\beta})$ is twice continuously differentiable with respect to \mathbf{s} , and $\boldsymbol{\alpha}(\mathbf{s})$ is twice continuously differentiable. Also in the expression

$$R(a(\cdot), \mathbf{b}(\cdot), \boldsymbol{\beta}) = E \left\{ \left(Y_t - a(\boldsymbol{\beta}^\tau \mathbf{X}_t) - \mathbf{b}(\boldsymbol{\beta}^\tau \mathbf{X}_t)^\tau \check{\check{\mathbf{X}}}_t \right)^2 w(\boldsymbol{\beta}^\tau \mathbf{X}_t) \right\},$$

differentiation with respect to β and the expectation are exchangeable.

(ii) The density function $f_{\beta^T \mathbf{X}_t(\mathbf{s})}(\mathbf{s}, z)$ of $\beta^T \mathbf{X}_t(\mathbf{s})$ is continuous. For any fixed $\mathbf{s} \in \mathcal{S}$, it is uniformly bounded away from 0 for $z \in [-L, L]$ and $\beta \in \mathbb{B}$, where $L > 0$ is a constant. Furthermore, the joint probability density function of $(\beta^T \mathbf{X}_{t_0}(\mathbf{s}), \beta^T \mathbf{X}_{t_1}(\mathbf{s}), \dots, \beta^T \mathbf{X}_{t_k}(\mathbf{s}))$ exists and is bounded uniformly for any $t_0 < t_1 < \dots < t_k$ and $0 \leq k \leq 2(r-1)$ and $\beta \in \mathbb{B}$, where $r > 3d$ is a positive integer.

(iii) Denote by $f_{Z,Z}(\mathbf{s}_1, \mathbf{s}_2; z_1, z_2)$ the joint density function of $Z_t(\mathbf{s}_1)$ and $Z_t(\mathbf{s}_2)$. Then $f_{Z,Z}(\mathbf{s}_1, \mathbf{s}_2; z, z)$ is continuous in z , and it converges to $f^*(\mathbf{s}, z)$ as $\|\mathbf{s}_i - \mathbf{s}\| \rightarrow 0$ for $i = 1, 2$.

C5. $E|Y_t(\mathbf{s})|^{\varrho r} < \infty$ and $E\|\mathbf{X}_t(\mathbf{s})\|^{\varrho r} < \infty$. Furthermore, it holds for some $\varrho > 6$ and r given in C4 that $\sup_{\beta \in \mathbb{B}} E|Y_t(\mathbf{s}) - \mathbf{g}(\mathbf{s}, \beta^T \mathbf{X}_t(\mathbf{s}), \beta)^T \mathbb{X}_t|^{\varrho r} < \infty$.

C6. Matrix $E(\mathbb{X}_t \mathbb{X}_t^T | \beta^T \mathbf{X}_t = z)$ is positively definite for $z \in [-L, L]$ and $\beta \in \mathbb{B}$.

C7. For each fixed $\mathbf{s} \in \mathcal{S}$, the time series $\{(Y_t(\mathbf{s}), \mathbf{X}_t(\mathbf{s})), t \geq 1\}$ is strictly stationary and β -mixing with the mixing coefficients satisfying the condition $\beta(t) = O(t^{-b})$ for some $b > \max\{2(\varrho r + 1)/(\varrho r - 2), (r + a)/(1 - 2/\varrho)\}$, where r and ϱ are specified in (C4) and (C5), and $a \geq (r\varrho - 2)r/(2 + r\varrho - 4r)$.

C8. There exists a continuous sampling intensity function (i.e. density function) f defined on \mathcal{S} for which $N^{-1} \sum_{i=1}^N I(\mathbf{s}_i \in A) \rightarrow \int_A f(\mathbf{s}) d\mathbf{s}$ for any measurable set $A \subset \mathcal{S}$.

C9. $W(\cdot)$ is a symmetric density function on \mathbb{R}^2 with bounded support. $K(\cdot)$ is a bounded and symmetric density function on \mathbb{R}^1 with bounded support S_K . Furthermore, $|K(x) - K(y)| \leq C|x - y|$ for $x, y \in S_K$ and some $0 < C < \infty$.

C10. As $N \rightarrow \infty$, $\tilde{h} \rightarrow 0$ and $N\tilde{h}^2 \rightarrow \infty$. As $T \rightarrow \infty$, $Th^4 = O(1)$, $Th^{3+3d/r} \rightarrow \infty$, and $\liminf_{T \rightarrow \infty} Th^{\frac{2(r-1)a + (\varrho r - 2)}{(a+1)\varrho}} > 0$ for r given in C4, ϱ given in C5 and a, b given in C7. Furthermore, there exists a sequence of positive integers $m_T \rightarrow \infty$ such that $m_T = o((Th)^{1/2})$, $Tm_T^{-b} \rightarrow 0$ and $m_T h^{\frac{2(\varrho r - 2)}{2+b(\varrho r - 2)}} > 1$.

Conditions C1 and C2 were introduced in Zhang *et al.* (2003). C2 manifests the so-called nugget effect in the noise process when $\sigma_2^2(\mathbf{s}_1) > 0$. The concept of nugget effect was introduced by G. Matheron in early 1960's. It implies that the variogram $E\{\varepsilon_t(\mathbf{s}_1) - \varepsilon_t(\mathbf{s}_2)\}^2$ does not converge to 0 as $\|\mathbf{s}_1 - \mathbf{s}_2\| \rightarrow 0$, or equivalently, the function $\tilde{\gamma}(\mathbf{s}) \equiv \gamma(\mathbf{s}_1 + \mathbf{s}, \mathbf{s}_1)$ is not continuous at $\mathbf{s} = 0$ for any given $\mathbf{s}_1 \in \mathcal{S}$. Note that $\varepsilon_{1,t}(\mathbf{s})$ in (6.2) represents so-called system noise which

typically has continuous sample realisation (in \mathbf{s}), while $\varepsilon_{2,t}(\mathbf{s})$ stands for microscale variation and/or measurement noise; see Cressie (1993, section 2.3.1). Condition C3 represents the possible nugget effect in the regressor process $\mathbf{X}_t(\mathbf{s})$. Note that when $\mathbf{X}_t(\mathbf{s})$ contains some lagged values of $Y_t(\mathbf{s})$, C3 may be implied by C2.

The function $f^*(\mathbf{s}, \cdot)$ in C4(iii) may not be the density function of $Z_t(\mathbf{s})$ due to the nugget effect; see the discussion below (3.1) in Lu *et al.* (2008). Other smooth conditions in C4 and the moment conditions in C5 are standard, though some of them may be reduced further at the cost of an increase of the already lengthy technical arguments. C6 is not essential and is introduced for technical convenience.

C7 imposes the β -mixing condition on the time series. The complex restriction on the mixing coefficients is required to ensure that the time series estimators $\hat{\alpha}(\mathbf{s})$, $\hat{a}(\mathbf{s})$ and $\hat{\mathbf{b}}(\mathbf{s})$ converge uniformly in \mathbf{s} . Further discussion on mixing time series may be found in, for example, Section 2.6 of Fan and Yao (2003).

C8 specifies that the asymptotic approximations in the spatial domain are derived from the fixed-domain or infill asymptotics (Cressie, 1993, Section 3.3).

The conditions on W and K in C9 and \tilde{h} in C10 are standard for nonparametric regression. More sophisticated conditions on the bandwidth h in C10 are required to ensure the uniform convergence of the time series estimators.

Acknowledgment. The authors thank Joint Editor, Professor George Casella, an Associate Editor and two referees for very helpful comments and suggestions.

References

- Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic: Dordrecht, Netherlands.
- Biau, G. and Cadre, B. (2004). Nonparametric spatial prediction. *Stat. Inference Stoch. Process.* **7**, 327-349.
- Brunsdon, C., McClatchey, J. and Unwin, D.J. (2001). Spatial variations in the average rainfall-altitude relationship in Great Britain: an approach using geographically weighted regression. *International Journal of Climatology*, **21**, 455-466.
- Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York.
- Cressie, N.A.C. (1993). *Statistics For Spatial Data*. Wiley, New York.

- Diggle, P.J. (2003). *Statistical analysis of spatial point patterns*. Arnold, London.
- Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.
- Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society*, **B**, **65**, 57-80.
- Fotheringham, A.S., Charlton, M.E. and Brunsdon, C. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and planning A*, **30**, 1905-1927.
- Fuentes M. (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetrics* 12: 469-483.
- Gao, J., Lu, Z. and Tjøstheim, D. (2006). Estimation in semiparametric spatial regression. *Annals of Statistics*, **34**, 1395-1435.
- Guyon, X. (1995). *Random Fields on a Network: Modelling, Statistics, and Application*. Springer-Verlag, New York.
- Hall, P. and Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection (with discussions). *Journal of the Royal Statistical Society*, **B**, **54**, 475-530.
- Hallin, M., Lu, Z. and Tran, L. T. (2004a). Density estimation for spatial processes: the L_1 theory. *Journal of Multivariate Analysis*, **88**, 61-75.
- Hallin, M., Lu, Z. and Tran, L. T. (2004b). Local Linear Spatial Regression. *Annals of Statistics*, **32**, 2469-2500.
- Härdle, W. and Vieu, P. (1992) Kernel regression smoothing of time series. *Journal of Time Series Analysis*, **13**, 209-224.
- Jones, M. C. (1991). The roles of ISE and MISE in density estimation. *Statistics and Probability Letters*, **12**, 51-56.
- Kalnay E., M. Kanamitsu, R. Kitsler, W. Collins, D. Deaven, L. Candin, M. Iredell, S. Saha, G. White, J. Wollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph, 1996: The NCEP/NCAR 40-year Reanalysis Project. *Bulletin of the American Meteorological Society*. **77**(3), 437-471.
- Kim, T. Y. and Cox, D. D. (1995). Asymptotic behaviors of some measures of accuracy in non-parametric curve estimation with dependent observations. *Journal of multivariate analysis*, **53**, 67-93.
- Knorr-Held, L. (2000). Bayesian modelling of inseparable space time variation in disease risk. *Statistics in Medicine*, **19**, 2555-567.
- Lagazio, C., Dreassi, E. and Biggeri, A. (2001). A hierarchical Bayesian model for space time variation of disease risk. *Statistical Modelling* **1**, 17-19.
- Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Ann. Statist.*, **27**, 415-438.

- Lu Z., Lundervold, A., Tjøstheim, D. and Yao, Q. (2007). Exploring spatial nonlinearity using additive approximation. *Bernoulli* **13**, 447-472.
- Lu, Z., Tjøstheim, D. and Yao, Q. (2007). Adaptive Varying-Coefficient Linear Models for Stochastic Processes: Asymptotic Theory. *Statistica Sinica*, **17**, 177-197 . [Additional proofs for “Adaptive varying-coefficient linear models for stochastic processes: asymptotic theory”. Available at the website of *Statistica Sinica*.]
- Lu, Z., Tjøstheim, D. and Yao, Q. (2008). Spatial smoothing, Nugget effect and infill asymptotics. *Statistics and Probability Letters*, **78**, 3145–3151.
- Mammen, E. (1990). A short note on optimal bandwidth selection for kernel estimators. *Statistics and Probability Letters*, **9**, 23–25.
- Matheron, G. (1976). A simple substitute for conditional expectation: the disjunctive kriging. *Proceedings, NATO ASI: Advanced Geostatistics in the Mining Industry* (Rome, Oct. 1975), 221-236.
- Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in C++: The art of scientific computing*. Cambridge University Press: Cambridge.
- Quintela-del-Río, A. (1996). Comparison of bandwidth selectors in nonparametric regression under dependence, *Comput. Statist. Data Anal.* **21**, 563–580.
- Ripley, B.D. (1981). *Spatial Statistics*. John Wiley & Sons: New York.
- Rivoirard, J. (1994). *Introduction to Disjunctive Kriging and Non-linear Geostatistics*. Clarendon Press, Oxford.
- Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257–1270.
- Rupert, D. (1997). Empirical-bias bandwidths for local polynomial regression and density estimation. *Journal of the American Statistical Association*, **92**, 1049-1062.
- Stein, M.L. (1999). *Interpolation of Spatial Data*. Springer, New York.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, B*, **36**. 111-147.
- Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association*, **94**, 1275-1285.
- Xia, Y. and Li, W. K. (2002). Asymptotic Behavior of Bandwidth Selected by the Cross-Validation Method for Local Polynomial Fitting. *Journal of multivariate analysis*, **83**, 265-287.
- Yao, Q. and Brockwell, P.J. (2006). Gaussian maximum likelihood estimation for ARMA models II: spatial processes. *Bernoulli*, **12**, 403-429.
- Zhang, W., Yao, Q., Tong, H. and Stenseth, N.C. (2003). Smoothing for spatial-temporal models and its application in modelling muskrat-mink interaction. *Biometrics*, **59**, 813-821.