UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Center For International Economics

Working Paper Series

Working Paper No. 2026-03

A semiparametric spatial FARIMA applied in the presence of spatial seasonality

Dominik Schulz, Thi Thu Huong Do and Yuanhua Feng

März 2026

Center For International Economics

A semiparametric spatial FARIMA applied in the presence of spatial seasonality

Dominik Schulz, Thi Thu Huong Do and Yuanhua Feng
Faculty of Business Administration and Economics, Paderborn University

March 25, 2026

Abstract

A semiparametric spatial long-memory time series model, called Semi-SFARIMA, in the presence of a seasonality surface and random intraday effects is introduced and defined. Initially, a local polynomial double conditional smoothing (LP-DCS) technique is employed to smooth the seasonal surface in the nonparametric part. Subsequently, intraday curves are identified using automated local polynomial smoothing in a one-dimensional context. Ultimately, residual series obtained from the preceding steps undergo modeling using a (parametric) spatial FARIMA framework. The statistical significance of the estimated spatial FARIMA parameters are assessed through asymptotically valid standard errors. The estimation method is illustrated by in-depth analysis of air temperature from two distinct weather stations, Yuma and Murphy. Since the example series include missing values, an imputation method for univariate time series is considered as a preliminary treatment during the data collection and processing. The proposed ideas prove to be useful in practice when being confronted with spatial time series with two-directional seasonality patterns.

Keywords: Spatial time series, Semi-SFARIMA, seasonality surface, imputation

JEL Codes: C14, C23, C51

1 Introduction

Numerous studies in the literature suggest that various time series data, such as revenues series, inflation rate, monetary aggregates, squared returns, temperature, and air pollution data, often display long-memory characteristics and possible cyclical behaviour (Ding et al., 1993; Granger and Ding, 1996; Andersen and Bollerslev, 1997; Baillie and Chung, 2002; Beran et al., 2015). As a result, various statistical approaches have been suggested for modeling such data in the univariate scenario. These include Gegenbauer autoregressive moving-average processes (GARMA) (Andel, 1986; Gray et al., 1989; Gray et al., 1994), the seasonal autoregressive fractionally integrated moving-average (SARFIMA) model (Porter-Hudak, 1990), k-factor GARMA processes (Giraitis and Leipus, 1995; Woodward et al., 1998), and flexible seasonal fractionally integrated processes (flexible ARFISMA, Hassler, 1994). Generally, in some fields such as biology, ethology, epidemiology, or bio-geography, and especially in climatology, spatial dependence usually indicates a scenario where the values observed at one location or region, such as observation i, are influenced by the values of neighboring observations in close proximity (Reynolds and Smith, 1994; Snell et al., 2000; Lin et al., 2002; Haining, 2003; Hubbard and You, 2005; Stahl et al., 2006; You et al., 2008; Wang et al., 2009, among others). Adopting this feature, when high-frequency data replaces observations of regions or locations, a spatial time series emerges with two dimensions of time. Quite often, but not necessarily, one dimension corresponds to the interday, i.e. the day to day, dimension, while the other pertains to the intraday dimensions, i.e. it contains observations within a given day. When observing such a spatial series, for example environmental data, an analyst could be confronted with a two-fold seasonality: firstly, there could be a pattern that repeats within each given day, while secondly, another pattern could arise at each given time of day over multiple days. A time series model with such a spatial seasonality and long-memory errors is at the center of this article, which is, to our knowledge, not represented in existing literature.

The remaining part of this paper is structured as follows. Section 2 proceeds with the definition of a Semi-SFARIMA model incorporating a smooth function for potential seasonal estimation adjustments under spatial FARIMA errors using the `DCSmooth` R-package (Schäfer, 2021a). Furthermore, this section also delves into the design and definition of spatial FARIMA models considered as the error process. Section 3.1 continues to introduce and explains briefly the double conditional smoothing (DCS) algorithm, including the procedure from kernel regression pioneered by Beran, 1994; Beran et al., 2015 to the use of local polynomial regression (Schäfer, 2021b; Letmathe et al., 2024). Next, Section 4 explains the estimation method for spatial seasonality which applies to the chosen examples in the application part. In Section 5, the methodology for estimating individual intraday curves is provided. Prior to Semi-SFARIMA modelling of real-world data using `DCSmooth`, Section 6 addresses a selection of missing value imputation methods, as the selected example series exhibit gaps in rare cases. Section 7 then presents some findings from the analysis of two air temperature datasets from two weather stations, one in Yuma (Arizona) and the other in Murphy (Idaho), comprising over 631,000 observations spanning six consecutive years from 2015 to 2020 with 288 observations per day at the frequency of 5 minutes. Finally, Section 8 concludes and provides some suggestions for further research.

2 Definition of the Semi-SFARIMA

By default, it is assumed that the data follow a Semi-SFARIMA (Peitz and Feng, 2015; Schäfer, 2021b). The setup of the two-dimensional, spatial model is

$$y_{a,b} = m\left(x_{1,a}, x_{2,b}\right) + \tau_{a,b}, \tag{1}$$

where $y_{a,b} = y\left(x_{1,a}, x_{2,b}\right)$ is the observed spatial time series; $m(\bullet,\bullet)$ is considered to be a smooth spatial seasonality surface, following a nonparametric regression function, and $\tau_{a,b}$ are stationary spatial errors with long memory, zero-mean and a finite variance $\mathrm{var}(\tau_{a,b}) = \sigma_\tau^2$. Accordingly, the random field $y_{a,b}$ of the Model (1) forms a triangular array with the sets of design points $X_1 = \{x_{1,a}\}, a = 1, ..., n_1$, and $X_2 = \{x_{2,b}\}, b = 1, ..., n_2$. The points $x_{1,a}$, $x_{2,b}$ are given by the rescaled variables $x_{1,a} = a/n_1$ and $x_{2,b} = b/n_2$ on the range [0,1] $\times$ [0,1]. Please note that outside of the context of this article, $m(x_{1,a}, x_{2,b})$ is in more general terms an (unconditional) mean surface in $y_{a,b}$, if $y_{a,b}$ is a Semi-SFARIMA. By this definition, the term semiparametric spatial FARIMA involves the estimation of two parts: the estimation of m as the nonparametric part and the employment of a parametric model (specifically, spatial FARIMA process) for modeling the error structure. In this paper, following the previous definition given in (1), exclusively equidistant series in both time dimensions are considered. In particular, note that we consider dimension a to be the interday dimension, i.e. the time dimension for the days, and b the intraday dimension, i.e. the time dimension for the time of day, throughout in context of our own empirical analyses in Section 7.

Model (1) is an adequate choice, if the data is roughly stationary after removing a suitable estimate of the entire surface $m(\cdot,\cdot)$. Nonetheless, since we find random and strongly varying intraday curves in Section 7 after removing estimated surfaces, an extension of Model (1) is motivated, where we account for these random effects. Denote by $g_a(x_{2,b})$ a random curve on day a over the day time $x_{2,b}$ with $E\left[g_a(\cdot)\right] \equiv 0$, where the expectation is over all observation days. Hence why the proposed extension of Model (1) is

$$y_{a,b} = m(x_{1,a}, x_{2,b}) + g_a(x_{2,b}) + \tau_{a,b}. \tag{2}$$

We call this model a *Semi-SFARIMA with random effects*. Note that $r_{a,b} \equiv g_a(x_{2,b}) + \tau_{a,b}$. In this context, both $m(\cdot,\cdot)$ and the collection of curves $(g_a(\cdot))$, $a = 1\ldots, n_1$, need to be estimated before analyzing the resulting residuals through a parameteric framework. The estimation steps for the components $m(\cdot,\cdot)$ and $g_a(\cdot)$ for Model (2) will be thoroughly elucidated in Section 4. This idea of random intraday curves is closely related to that by Feng and Peitz (2013), who proposed a semiparametric spatial volatility model with individual random effects both for the intraday and the interday dimension.

In general, the process $\tau_{a,b}$ is referred to as an SFARIMA(p_1, p_2)(q_1, q_2) process (Beran et al., 2009), d_1 and d_2 are implied, which is a spatial extension of the univariate FARIMA model (Granger and Joyeux, 1980; Hosking, 1981):

$$\phi_1\left(B_1\right)\phi_2\left(B_2\right)\left(1-B_1\right)^{d_1}\left(1-B_2\right)^{d_2}\tau_{a,b} = \theta_1\left(B_1\right)\theta_2\left(B_2\right)\upsilon_{a,b}, \tag{3}$$

where $\upsilon_{a,b}$ represents independent and identically distributed (iid) random variables with mean zero and a finite variance σ_υ^2. Moreover, $B = (B_1, B_2)$ signifies the backshift operators in the two dimensions. If $\xi\left(B\right) = \xi_1\left(B_1\right)\xi_2\left(B_2\right)$ and $\psi\left(B\right) = \psi_1\left(B_1\right)\psi_2\left(B_2\right)$, where

$$\xi_k\left(B_k\right) = \left(1-B_k\right)^{d_k}\phi_k\left(B_k\right)\theta_k^{-1}\left(B_k\right) = 1 + \sum_{i=1}^{\infty}\xi_{k,i}B_k^i, \tag{4}$$

$$\psi_k\left(B_k\right) = \left(1-B_k\right)^{-d_k}\phi_k^{-1}\left(B_k\right)\theta_k\left(B_k\right) = 1 + \sum_{j=1}^{\infty}\psi_{k,j}B_k^j, \tag{5}$$

$k = 1, 2$ and where $\xi_{k,0} = \psi_{k,0} = 1$, then, if $\tau_{a,b}$ is stationary and invertible, Equation (3) is equivalent to

$$\upsilon_{a,b} = \xi_1 (B_1) \xi_2 (B_2) \tau_{a,b} = \xi (B) \tau_{a,b} \qquad \text{and} \tag{6}$$

$$\tau_{a,b} = \psi_1 (B_1) \psi_2 (B_2) \upsilon_{a,b} = \psi (B) \upsilon_{a,b}. \tag{7}$$

Equation (6) and (7) are the spatial infinite-order AR and the spatial infinite-order MA representations, respectively. It is known that $0 \leq d_1, d_2 < 0.5$ denotes two memory parameters, while $\phi_1 (B_1)$ and $\phi_2 (B_2)$ stand for the AR-polynomials, and $\theta_1 (B_1)$ and $\theta_2 (B_2)$ denote the MA-polynomials of marginal time series models in each dimension. It is customary to note that these polynomials are without common factors and that they possess all roots located outside the unit circle and follow

$$\phi_k (B_k) = 1 - \phi_{k,1} B_k - \cdots - \phi_{k,p_k} B_k^{p_k} = 1 - \sum_{i=1}^{p_k} \phi_{k,i} B_k^i \qquad \text{and} \tag{8}$$

$$\theta_k (B_k) = 1 + \theta_{k,1} B_k + \cdots + \theta_{k,q_k} B_k^{q_k} = 1 + \sum_{j=1}^{q_k} \theta_{k,j} B_k^j \tag{9}$$

with $k = 1, 2$, where all ϕ_{k_i} and all $\theta_{k,j}$ are real-valued coefficients. In such cases, the spatial process corresponds to the product of two univariate FARIMA processes. Moreover, within a separable SFARIMA process, parameter estimation can be streamlined to involve the estimation of two univariate FARIMA processes in two directions (Martin, 1979), akin to the local polynomial double conditional smoothing (LP-DCS) procedure. Generally, the techniques utilized for parameter estimation in univariate FARIMA models can be readily adapted to estimate the parameters in the polynomials $\phi_1 (B_1)$, $\phi_2 (B_2)$, $\theta_1 (B_1)$, $\theta_2 (B_2)$ and σ_υ^2 in the SFARIMA model outlined in Equation (3).

3 Estimation of Semi-SFARIMA

Based on Model (1), we propose to estimate the smooth surface $m(x_{1,a}, x_{2,b})$ and the SFARIMA $\tau_{a,b}$ in a two-step procedure using LP-DCS and a least squares estimator of the SFARIMA following the descriptions in Feng et al. (2021), however without automatic bandwidth selection. Therefore, in the subsequent Sections 3.1 and 3.2, respectively, brief summaries on these topics are provided. The exact LP-DCS-based estimation approach for $m(x_{1,a}, x_{2,b})$, if $m(x_{1,a}, x_{2,b})$ is a spatial seasonality surface, is described in Section 4.

3.1 Summary of double conditional smoothing

Generally, smoothing is understood as the estimation of the mean surface of the observations, or functional time series by traditional bi-variate (2D) nonparametric regression approaches in literature (Müller, 1988; Ruppert and Wand, 1994; Härdle and Müller, 2000; Scott, 2015). The innovative DCS represents a significant advancement in the existing literature, markedly enhancing computational efficiency (Peitz and Feng, 2015; Schäfer, 2021b). This improvement is particularly notable because traditional 2D smoothing techniques become slow as the number of observations in both dimensions increase. In summary, the development of the DCS method has progressed through two distinct phases. Initially, it involved bi-variate kernel regression using a bi-variate product kernel estimator and kernel weights (Müller, 1988; Müller and Prewitt, 1993). This was followed by the LP-DCS with the incorporation of boundary correction to address boundary effects (Beran and Feng, 2002b; Schäfer, 2021b; Letmathe et al., 2024). The method was refined to enhance computational efficiency, resulting in a faster functional smoothing scheme designed to eliminate redundant computations and reduce the runtime of the IPI algorithm implementation.

As per knowledge, local polynomial regression for multivariate data investigated by some authors, e.g., Ruppert and Wand (1994), Yang and Tschernig (1999), Hallin et al. (2004) and Wang and Wang (2009) together with the DCS technique proposed by Peitz and Feng (2015), therein however under kernel regression and iid errors, are the foundations for the `DCSmooth` package by Schäfer (2021a) for the statistical software R. Furthermore, the combination of DCS and local polynomial regression for multivariate data solves the boundary problem of kernel regression, in which the order of the bias at the boundary differs from the bias in the interior. This becomes particularly challenging when managing two-dimensional data, given that the ratio of the boundary region to the interior typically exceeds that observed in the univariate scenario. In literature, some authors proposed to correct this problem by using boundary correction kernels (Gasser and Müller, 1979; Müller, 1992; Müller and Wang, 1994), yet, estimation at the boundary is still unstable, especially for the estimation of derivatives. Hence, the introduction of local polynomial regression instead of kernel regression, which has an automatic boundary correction, mitigated that problem (Fan and Gijbels, 1992; Fan and Gijbels, 1996). Under the assumption that m is at least $(o_1 + 1, o_2 + 1)$-times differentiable at a point

$(x_{1,0}, x_{2,0})$, $m\left(x_{1,a}, x_{2,b}\right)$ in (1) can be approximated by a local polynomial of order (o_1, o_2) for $(x_{1,a}, x_{2,b})$ in a neighbourhood of $(x_{1,0}, x_{2,0})$. Then the overall minimization problem can be stated to be

$$\begin{aligned} Q(x_{1,0}, x_{2,0}) = &\sum_{a=1}^{n_1}\sum_{b=1}^{n_2}\left\{y_{a,b} - \sum_{s=0}^{o_1}\sum_{t=0}^{o_2}\left(x_{1,a} - x_{1,0}\right)^s\left(x_{2,b} - x_{2,0}\right)^t\beta_{s,t}(x_{1,0}, x_{2,0})\right\}^2 \\ &\times W\left(\frac{x_{1,a} - x_{1,0}}{h_1}, \frac{x_{2,b} - x_{2,0}}{h_2}\right) \qquad \Rightarrow \min, \end{aligned} \tag{10}$$

where $\beta_{s,t}(x_{1,0}, x_{2,0})$ are the coefficients of interest in the proximity of $(x_{1,0}, x_{2,0})$ so that the estimator

$$\hat{m}^{(\kappa_1,\kappa_2)}(x_{1,0}, x_{2,0}) = \kappa_1!\kappa_2!\hat{\beta}_{\kappa_1,\kappa_2}(x_{1,0}, x_{2,0}) \tag{11}$$

arises for the derivative of orders κ_1 and κ_2 for each respective dimension of $m(x_{1,0}, x_{2,0})$. Moreover, $W(u_1, u_2) = W_1(u_1)W_2(u_2)$ are bivariate product weights with W_1 and W_2 being weighting functions each with compact support $[-1, 1]$ that are allowed to have boundary modification. In practice, this single-step approach is however computationally extensive, which is why Schäfer (2021b, pp. 65–68) exemplifies DCS in context of two-dimensional local polynomial regression under selected assumptions.

Schäfer (2021b, pp. 65–68) gives

$$Q_{1|2}(x_{1,0}) = \sum_{a=1}^{n_1}\left\{y_{a,b} - \sum_{s=0}^{o_1}\beta_{1|2,s}(x_{1,0})\left(x_{1,a} - x_{1,0}\right)^s\right\}^2 W_1\left(\frac{x_{1,a} - x_{1,0}}{h_1}\right), \tag{12}$$

$b = 1, \ldots, n_2, \quad$ and

$$Q_{2|1}(x_{2,0}) = \sum_{b=1}^{n_2}\left\{y_{a,b} - \sum_{s=0}^{o_2}\beta_{2|1,s}(x_{2,0})\left(x_{2,b} - x_{2,0}\right)^s\right\}^2 W_2\left(\frac{x_{2,b} - x_{2,0}}{h_2}\right), \tag{13}$$

$a = 1, \ldots, n_1,$

as the intermediate weighted sums of squares, by conditioning on either $x_{1,a}$ or $x_{2,b}$, that need to be minimized, with solutions

$$\hat{m}^{(\kappa_1)}(x_{1,0} \mid x_{2,b}) = \kappa_1!\hat{\beta}_{1|2,\kappa_1}(x_{1,0}) \quad \text{and} \tag{14}$$

$$\hat{m}^{(\kappa_2)}(x_{2,0} \mid x_{1,a}) = \kappa_2!\hat{\beta}_{2|1,\kappa_2}(x_{2,0}). \tag{15}$$

Given these intermediate results, the equivalent solutions to the weighted sums of squares

$$Q_1(x_{1,0}) = \sum_{a=1}^{n_1}\left\{\hat{m}^{(\kappa_2)}(x_{2,0} \mid x_{1,a}) - \sum_{s=0}^{o_1}\beta_{1,s}(x_{1,0})\left(x_{1,a} - x_{1,0}\right)^s\right\}^2 W_1\left(\frac{x_{1,a} - x_{1,0}}{h_1}\right) \tag{16}$$

as well as

$$Q_2(x_{2,0}) = \sum_{b=1}^{n_2} \left\{ \hat{m}^{(\kappa_1)}(x_{1,0} \mid x_{2,b}) - \sum_{s=0}^{o_2} \beta_{2,s}(x_{2,0}) \left(x_{2,b} - x_{2,0}\right)^s \right\}^2 W_2 \left(\frac{x_{2,b} - x_{2,0}}{h_2} \right) \quad (17)$$

are either

$$\hat{m}^{(\kappa_1,\kappa_2)}(x_{1,0}, x_{2,0}) = \kappa_1! \hat{\beta}_{1,\kappa_1}(x_{1,0}) \quad \text{or} \quad (18)$$

$$\hat{m}^{(\kappa_1,\kappa_2)}(x_{1,0}, x_{2,0}) = \kappa_2! \hat{\beta}_{2,\kappa_2}(x_{2,0}), \quad (19)$$

Thus, DCS for local polynomial regression is a two-step procedure, where at first univariate local polynomial smoothing is done for each subseries in one dimension and where in a second step the intermediate results are considered for smoothing in the other dimension (Schäfer, 2021b, pp. 65–68). For each of the problems (12) to (17) there exists a closed-form solution. Denote by

$$\mathbf{X}_k = \begin{pmatrix} 1 & (x_{k,1} - x_{k,0}) & (x_{k,1} - x_{k,0})^2 & \dots & (x_{k,1} - x_{k,0})^{o_k} \\ 1 & (x_{k,2} - x_{k,0}) & (x_{k,2} - x_{k,0})^2 & \dots & (x_{k,2} - x_{k,0})^{o_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (x_{k,n_k} - x_{k,0}) & (x_{k,n_k} - x_{k,0})^2 & \dots & (x_{k,n_k} - x_{k,0})^{o_k} \end{pmatrix}, \quad k = 1, 2, \quad (20)$$

the $n_k \times (o_k + 1)$ design matrix. Furthermore, let $\mathbf{z} = (z_1, z_2, \dots, z_{n_k})^{\mathrm{T}}$ be a $n_k \times 1$ vector with the values of the (one-dimensional) regressand, for example with the elements of $y_{a,b}$ either given a or b, of $\hat{m}^{(\kappa_2)}(x_{2,0} \mid x_{1,a})$ or of $\hat{m}^{(\kappa_1)}(x_{1,0} \mid x_{2,b})$. In addition, $\mathbf{W}$ must be a $n_k \times n_k$ diagonal matrix containing the weights following the weighting function on the main diagnoal. Then,

$$\hat{l}(x_{k,0}) = \kappa_k! \iota_{\kappa_k+1}^{\mathrm{T}} \left(\mathbf{X}_k^{\mathrm{T}} \mathbf{W} \mathbf{X}_k\right)^{-1} \mathbf{X}_k^{\mathrm{T}} \mathbf{z}, \quad k = 1, 2, \quad (21)$$

where ι_i is a $(o_k + 1) \times 1$ vector with zeros everywhere, with exception of the i-th element, which is one. $\hat{l}(x_{k,0})$ is the (univariate) local polynomial estimator of the quantity of interest at the time point $x_{k,0}$.

Regarding boundary modification within LP-DCS, the boundary region can be delineated as encompassing all points within X_1 and X_2 that lie within a distance no greater than the bandwidths from the boundaries at 0 or 1 (Schäfer, 2021b). As highlighted by Schäfer (2021b), there is an increasing share of the boundary region in the total area of support [0,1] $\times$ [0,1] in two-dimensional models in comparison to one-dimensional ones.

A major drawback of (bivariate) local polynomial regression at boundary points is the potential discontinuity in the bias of the estimator between interior and boundary points (Schäfer, 2021b). To tackle this problem, besides the truncated weight function used in the interior, Schäfer (2021b) proposes to consider two classes of boundary-modified local regression weights (23) and (24), adapting the ideas of boundary kernels from Müller (1992) and Müller and Wang (1994) in order to continue to estimate the boundary regions where the truncated part of the weight function (22) for the interior region cannot. At a boundary point $x_{\cdot,a}$, for $q \in [0,1]$ and a smoothness parameter $\mu \in \{0,1,2,\dots\}$, for the right boundary, these weights are:

$$W_q^0(u) = (1+u)^\mu (1-u)^\mu, \quad u \in [-1,q], \tag{22}$$

$$W_q^r(u) = (1+u)^\mu (q-u)^\mu, \quad u \in [-1,q], \tag{23}$$

$$W_q^s(u) = (1+u)^\mu (q-u)^{\mu'}, \; u \in [-1,q], \; \mu' = \max(0, \mu-1). \tag{24}$$

The corresponding left boundary weighting functions have support $u \in [-q,1]$. Among these three methods, (24) provides the lowest bias (Schäfer, 2021b). Thus, in our analysis in Section 7 using the R package `DCSmooth`, we will use the default of Müller Wang kernels (Müller and Wang, 1994) of orders $(c,\mu,\kappa) = (2,2,0)$ in both dimensions, where c is the kernel order and κ is the order of derivative (either κ_1 or κ_2), which are the default in the package and which correspond to a special case of (24).

Assume $b_1 = O(n_1^{\zeta_1})$ and $b_2 = O(n_2^{\zeta_2})$, where $\zeta_1, \zeta_2 > -1/c$ and furthermore presume both $(n_1 b_1)^{1-2d_1} b_1^{2\kappa_1} \to \infty$ and $(n_2 b_2)^{1-2d_2} b_2^{2\kappa_2} \to \infty$ for $n_1 \to \infty$ and $n_2 \to \infty$, respectively (Feng et al., 2021). Furthermore, $m^{(\kappa_1,\kappa_2)}(x_{1,a}, x_{2,b})$ must be at least two times continously differentiable and Lipschitz continuous, while it is also assumed that $m^{(\kappa_1+1,\kappa_2+1)}(x_{1,a}, x_{2,b})$ is Lipschitz continuous. According to Theorem 3 in Feng et al. (2021), and given $c = o_1 + 1 + \kappa_2 = o_2 + 1 + \kappa_1$, $\lambda_1 = c - \kappa_2$, $\lambda_2 = \kappa_1$, $\alpha_k = (-1)^c \int_{-1}^1 \int_{-1}^1 u_1^{\lambda_k} u_2^{c-\lambda_k} K_{(\kappa_1,\kappa_2)}(u_1,u_2) \mathrm{d}u_1 \mathrm{d}u_2 / [\lambda_k! \, (c-\lambda_k)!]$ (for $k=1,2$), $\omega = c - \kappa_1 - \kappa_2$, $C = 4\sigma_v^2 c_{f_1} c_{f_2} \times R_1(d_1) R_2(d_2)$, and

$$R_k(d_k) = \Gamma(1-2d_k) \sin(\pi d_k) \int_{-1}^1 \int_{-1}^1 K_{k,\kappa_k}(u_1) K_{k,\kappa_k}(u_2) |u_1 - u_2|^{2d_k - 1} \mathrm{d}u_1 \mathrm{d}u_2, \tag{25}$$

$k = 1,2$, the asymptotic point-wise bias of $\hat{m}^{\kappa_1,\kappa_2}(x_{1,a}, x_{2,b})$ is

$$\begin{aligned} E\left(\hat{m}^{(\kappa_1,\kappa_2)}(x_{1,a}, x_{2,b}) - m^{(\kappa_1,\kappa_2)}(x_{1,a}, x_{2,b})\right) &= m^{(\lambda_1, c-\lambda_1)}(x_{1,a}, x_{2,b}) \left[\alpha_1 + o(1)\right] b_1^\delta \\ &\quad + m^{(\lambda_2, c-\lambda_2)}(x_{1,a}, x_{2,b}) \left[\alpha_2 + o(1)\right] b_2^\delta \end{aligned} \tag{26}$$

where $\delta = c - \kappa_1 - \kappa_2$, and its asymptotic point-wise variance is

$$\text{var}\left(\hat{m}^{\kappa_1,\kappa_2}(x_{1,a}, x_{2,b})\right) = C\left[(n_1b_1)^{1-2d_1}(n_2b_2)^{1-2d_2}b_1^{2\kappa_1}b_2^{2\kappa_2}\right]^{-1}[1+o(1)] \tag{27}$$

at an interior point. $K_{(\kappa_1,\kappa_2)}$ and K_{k,κ_k}, $k = 1, 2$, are asymptotically equivalent kernel functions. c_{f_1} and c_{f_2} are the sums of autocovariances (each divided by 2π) of ARMA processes with AR-polynomial $\phi(B) = \phi_1(B)$ or $\phi(B) = \phi_2(B)$ and MA-polynomial $\theta(B) = \theta_1(B)$ or $\theta(B) = \theta_2(B)$, correspondingly.

In terms of selecting optimal bandwidths, the minimizer of the asymptotic mean integrated squared error (AMISE) is commonly used as a measure of the goodness-of-fit (Herrmann et al., 1995). This bandwidth selection idea is also incorporated by Schäfer (2021b) in the context of semiparametric extensions of spatial ARIMA (SARIMA) and of SFARIMA models. In essence, AMISE is the (asymptotic integrated) variance plus the squared (asymptotic integrated) bias. For small bandwidths, there is small (squared) bias but large variance while the (squared) bias is large and variance is small for large bandwidths. Since there is a trade-off between (squared) bias and variance, AMISE is one option for bandwidth selection criterion in nonparametric regression. Within the DCSmooth package, bandwidth selection involves minimizing AMISE, with asymptotically valid formulas derived for the bandwidths. Within the package, a data-driven IPI procedure is employed to iteratively compute bandwidths until convergence is reached. The resulting bandwidths are the estimates of the asymptotically optimal bandwidths described in Schäfer (2021b). Nevertheless, practical implementation, exemplified in Section 7, reveals limitations when attempting to smooth the seasonality surface under spatial FARIMA errors with automatically chosen optimal bandwidths, as the bandwidth tends to converge to zero. Consequently, to address this scenario, manually selected bandwidths are employed during processing to ensure integration of potential long memory errors and the usage of the automated bandwidth selection by Schäfer (2021b) is omitted.

3.2 Summary of the least squares estimator

Regarding the SFARIMA process $\tau_{a,b}$, in (1), define $r_k = p_k + q_k$, $k = 1, 2$, and $r = r_1 + r_2$. Denote by $\delta_k = (d_k, \phi_{k,1}, \ldots, \phi_{k,p_k}, \theta_{k,1}, \ldots, \theta_{k,q_k})^T$ the $(r_k + 1) \times 1$ SFARIMA paramter vector for either dimension $k = 1, 2$, so that $\delta^0 = \left(\delta_1^{\mathrm{T}}, \delta_2^{\mathrm{T}}\right)^{\mathrm{T}}$ is a $(r + 2) \times 1$ vector.

According to Beran et al. (2009), a simple least squares estimator (LSE) for δ^0 is defined as follows: an SFARIMA process (3), if invertible, allows for the representation

$$\begin{aligned} v_{a,b} &= \xi(B)\tau_{a,b} \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\xi_{1,i}\xi_{2,j}\tau_{a-i,b-j} \end{aligned} \tag{28}$$

by making use of Equation (4). Consequently, given that $\tau_{a,b}$ are only available for $a = 1,\dots,n_1$ and $b = 1,\dots,n_2$, the innovations $v_{a,b}$ are approximated through

$$\hat{v}_{a,b}(\delta) = \sum_{i=0}^{a-1}\sum_{j=0}^{b-1}\xi_{1,i}\xi_{2,j}\tau_{a-i,b-j}, \tag{29}$$

which subsequently leads to the least squares minimization problem

$$\sum_{a=1}^{n_1}\sum_{b=1}^{n_2}\hat{v}_{a,b}^2(\delta) \Rightarrow \min. \tag{30}$$

The solution $\hat{\delta}$ to (30) is the LSE of δ^0, which does not have a closed-form solution and is therefore obtained numerically. Moreover, Feng et al. (2021) propose a computationally faster two-stage version of the LSE: firstly,

$$\hat{v}_{a,b}(\delta_2) = \sum_{j=0}^{b-1}\xi_{2,j}\tau_{a,b-j} \tag{31}$$

are calculated for all $1 \leq a \leq n_1$ and subsequently

$$\hat{v}_{a,b}(\delta) = \sum_{i=0}^{a-1}\xi_{1,i}\hat{v}_{a-i,b}(\delta_2). \tag{32}$$

for all $1 \leq b \leq n_2$. To find $\hat{\delta}$, once again criterion (30) is used.

Furthermore, under regulatory assumptions, Beran et al. (2009) provide (for a perfectly rectangular time lattice)

$$(n_1 \times n_2)^{\frac{1}{2}}\left(\hat{\delta} - \delta^0\right) \underset{d}{\to} Z \sim N\left(0, \mathbf{V}^{-1}\left(\delta^0\right)\right) \tag{33}$$

as the asymptotic distribution of the LSE $\hat{\delta}$, i.e. it is asymptotically normally distributed with mean vector δ^0 and with $(r+2) \times (r+2)$ variance-covariance matrix $\mathbf{C}_V = (n_1 \times n_2)^{-1}\mathbf{V}^{-1}(\delta^0)$, where

$$\mathbf{V}(\delta^0) = \begin{pmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{pmatrix}, \tag{34}$$

with $\mathbf{0}$ being zero matrices of adequate size and where $\mathbf{V}_1$ and $\mathbf{V}_2$ are defined through

$$\mathbf{V}_k = (4\pi)^{-1} \int_{-\pi}^{\pi} \left\{ \left[\frac{\partial \ln [f_k(\lambda)]}{\partial \delta_k} \right] \left[\frac{\partial \ln [f_k(\lambda)]}{\partial \delta_k} \right]^{\mathrm{T}} \mathrm{d}\lambda \right\}, \quad k = 1, 2. \tag{35}$$

$f_1(\lambda)$ and $f_2(\lambda)$ in Equation (35) are the marginal spectral densities of the SFARIMA given through

$$f_k(\lambda) = (2\pi)^{-1} \left|1 - e^{-i\lambda}\right|^{-2d_k} \left| \frac{\theta_k\left(e^{-i\lambda}\right)}{\phi_k\left(e^{-i\lambda}\right)} \right|^2, \quad k = 1, 2, \tag{36}$$

where $i = \sqrt{-1}$ and e is Euler's number (Beran et al., 2009), so that the spectral density of $\tau_{a,b}$ is given through $f(\lambda) = \sigma_v^2 f_1(\lambda) f_2(\lambda)$. As a consequence, we obtain

$$\begin{aligned} \ln [f_k(\lambda)] = & - \ln(2\pi) - d_k \ln [2 - 2\cos(\lambda)] \\ & + \ln \left\{ 1 + 2 \sum_{i=1}^{q_k} \theta_{k,i} \cos(i\lambda) + \sum_{i=1}^{q_k} \sum_{j=1}^{q_k} \theta_{k,i}\theta_{k,j} \cos[(i-j)\lambda] \right\} \\ & - \ln \left\{ 1 - 2 \sum_{i=1}^{p_k} \phi_{k,i} \cos(i\lambda) + \sum_{i=1}^{p_k} \sum_{j=1}^{p_k} \phi_{k,i}\phi_{k,j} \cos[(i-j)\lambda] \right\}, \end{aligned} \tag{37}$$

leading to the partial derivatives

$$\frac{\partial \ln [f_k(\lambda)]}{\partial \phi_{k,l}} = \frac{2\left\{\cos(l\lambda) - \sum_{i=1}^{p_k} \phi_{k,i} \cos[(l-i)\lambda]\right\}}{1 - 2\sum_{i=1}^{p_k} \phi_{k,i} \cos(i\lambda) + \sum_{i=1}^{p_k} \sum_{j=1}^{p_k} \phi_{k,i}\phi_{k,j} \cos[(i-j)\lambda]}, \tag{38}$$

$$\frac{\partial \ln [f_k(\lambda)]}{\partial \theta_{k,s}} = \frac{2\left\{\cos(s\lambda) + \sum_{i=1}^{q_k} \theta_{k,i} \cos[(s-i)\lambda]\right\}}{1 + 2\sum_{i=1}^{q_k} \theta_{k,i} \cos(i\lambda) + \sum_{i=1}^{q_k} \sum_{j=1}^{q_k} \theta_{k,i}\theta_{k,j} \cos[(i-j)\lambda]}, \quad \text{and} \tag{39}$$

$$\frac{\partial \ln [f_k(\lambda)]}{\partial d_k} = - \ln [2 - 2\cos(\lambda)], \tag{40}$$

$l = 1, \ldots, p_k$, $s = 1, \ldots, q_k$, all for $k = 1, 2$.

The result in Equation (33), while not being useful in small-sample settings, can be considered to obtain at least approximately valid standard errors of $\hat{\delta}$ under sufficiently large n_1 and n_2 by estimating $\sqrt{\mathrm{diag}\left(\mathbf{C}_V\right)}$ through plugging the elements of $\hat{\delta}$ into Equations (38), (39) and (40), in order to then obtain estimates of $\mathbf{V}_k$ in (35), $k = 1, 2$, through numerical integration. Estimating $\mathbf{V}\left(\delta^0\right)$ in (34) and subsequently $\mathbf{C}_V$ are then straightforward numerical computations. On a further note, Beran et al. (2009) describe that the statistic

$$S = (n_1 \times n_2)^{1/2} \frac{\hat{d}_1 - \hat{d}_2}{\sqrt{\sigma_1 + \sigma_2}}, \tag{41}$$

where $\hat{d}_k$ are the estimated d_k, $k = 1, 2$, and where σ_1 and σ_2 are the corresponding asymptotic standard deviations, is approximately standard normally distributed. Similar thoughts also hold for any other pair of estimators, one estimator from $\hat{\delta}_1$ and one from $\hat{\delta}_2$, due to the asymptotic uncorrelatedness between the two time dimensions implied by (34), which makes statistical tests on the equality of parameters for the two different time dimensions feasible.

Notwithstanding the previous elaborations, when fitting a Semi-SFARIMA model in practice, the LSE is not applied directly to a realization of $\tau_{a,b}$ but instead to residuals $\hat{\tau}_{a,b} = y_{a,b} - \hat{m}(x_{1,a}, x_{2,b}) - \hat{g}_a(x_{2,b})$ after preliminary steps, in which the smooth surface m and the intraday curves g_a in (1) are estimated via local polynomial smoothing techniques following the descriptions in Section 3.1. Beran and Feng (2002a) found for univariate SEMIFAR models that the effect of the errors in local linear or local cubic estimators on the subsequent estimators for the parametric part of the model is asymptotically negligible, as long as a bandwidth of the optimal order is considered. We assume, given that bandwidths are selected either manually for the surface or automatically for the intraday curves in our applications in Section 7, that approximately such optimal-order bandwidths are selected and that the result from Beran and Feng (2002a) can also be extended to the spatial case, so that the effect of the error in $\hat{m}(x_{1,a}, x_{2,b})$ on $\hat{\delta}$ is asymptotically negligible in our case and that moreover the findings on the asymptotic behavior of $\hat{\delta}$ stated in Beran et al. (2009) also hold here.

4 Estimation method for spatial seasonality

Assume that m in (1) is a spatial seasonal component with seasonal period s_1 regarding the first time dimension and seasonal period s_2 in the second time dimension. For the sake of simplified notation, we assume that $n_1 = s_1 \times m_1$ and $n_2 = s_2 \times m_2$, $m_1, m_2 \in \mathbb{N}_+$, are integer multiples of s_1 and s_2, however, the following ideas can also be easily adapted to scenarios, where this is not the case. The time points (a, b), $a = 1, \ldots, n_1$ and $b = 1, \ldots, n_2$, of $y_{a,b}$ in Model (2) are originally located on a grid $A = \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$, which can be divided into $N = n/(s_1 s_2) = m_1 \times m_2$ adjacent and non-overlapping sub-grids of equal size $s_1 \times s_2$, here denoted by $A_{i,j}$, $i = 1, \ldots, m_1$ and $j = 1, \ldots, m_2$, where

$A_{i,j} \subseteq A \; \forall i = 1, \dots, m_1$ and $j = 1, \dots, m_2$ and where

$$A_{i,j} = \{(a,b) : (i-1)s_1 + 1 \le a \le is_1, (j-1)s_2 + 1 \le b \le js_2\}\,. \tag{42}$$

Given the previous setting, we can, for each subperiod (i,j), define a submodel of (1) stated as

$$y^{[i,j]}_{a,b} = m^{[i,j]}(x^*_{1,a}, x^*_{2,b}) + g^{[i,j]}_a(x_{2,b}) + \tau^{[i,j]}_{a,b}, \quad (a,b) \in A_{i,j}, \tag{43}$$

where $[i,j]$, $i = 1, \dots, m_1$ and $j = 1, \dots, m_2$, is an indicator of the subperiod. Moreover, $x^*_{1,a} = a/s_1$ and $x^*_{2,b} = b/s_2$. We propose to estimate the seasonality surface $m^{[i,j]}(x^*_{1,a}, x^*_{2,b})$ in each subperiod based on Model (43), separately, using LP-DCS as discussed in Section 3.1. Denote the corresponding estimated surface in subperiod $[i,j]$ by $\tilde{m}^{[i,j]}(x^*_{1,a}, x^*_{2,b})$. In the following, two approaches to obtain the complete estimated seasonality surface $\hat{m}(x_{1,a}, x_{2,b})$ for Model (1) are suggested.

The first proposal is based on different sets of bandwidths $(h_{1,(i,j)}, h_{2,(i,j)})$ for each subperiod (i,j). Following this idea, $\tilde{m}^{[i,j]}(x^*_{1,a}, x^*_{2,b})$, as the estimated seasonality surface using LP-DCS in subperiod $[i,j]$, is obtained in each subperiod using a (potentially) different set of manually selected bandwidths. Noting that each pair (a,b) exists only once in the collection of subperiods $(A_{i,j})_{i=1,\dots,m_1,\, j=1,\dots,m_2}$, we consequently consider

$$\hat{m}(x_{1,a}, x_{2,b}) = \tilde{m}^{[i,j]}(x^*_{1,a}, x^*_{2,b}), \quad (i,j) = \operatorname{elem}\left(\{(u,v) \in M : (a,b) \in A_{u,v}\}\right), \tag{44}$$

where $M = \{1, \dots, m_1\} \times \{1, \dots, m_2\}$ and $\operatorname{elem}(\cdot)$ extracts the element of a single-element set, as the overall seasonality surface estimates. This method is denoted by **M1**. Since results by M1 are considered directly without further ensembling, common results regarding bias and variance for LP-DCS are unaffected at all estimation time points.

Secondly, consider the same set of bandwidths (h_1, h_2) throughout and obtain the quantities $\tilde{m}^{[i,j]}(x^*_{1,a}, x^*_{2,b})$ using this set of bandwidths based on Equation (43) using LP-DCS as discussed in Section 3.1 for each of the N sub-models. Then,

$$\hat{m}(x_{1,a}, x_{2,b}) = N^{-1} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \tilde{m}^{[i,j]}(x^*_{1,a^*(a,i)}, x^*_{2,b^*(b,j)}), \quad (a,b) \in A, \tag{45}$$

where $a^*(a,i) = (i-1)s_1 + [(a-1) \bmod s_1] + 1$ and $b^*(b,j) = (j-1)s_2 + [(b-1) \bmod s_2] + 1$. Equation (45) describes that for $\hat{m}$ on each subgrid $A_{i,j}$ an estimated average surface obtained from averaging over $\tilde{m}^{[i,j]}$ over all time grids $A_{i,j}$, $i = 1, \dots, m_1$ and $j = 1, \dots, m_2$, is repeated. This method will be called **M2** throughout this article. Since $\hat{m}$ obtained by this method is simply an arithmetic mean of multiple LP-DCS estimators (with the same settings) using the same set of bandwidths, the asymptotic bias (26) of $\hat{m}$ remains unaffected. The asymptotic variance is left open for further investigation.

When applying **M2**, we may encounter various problems due to the number of observations per sub-period not always being equal under certain circumstances. First and foremost, when we consider the seasonal cycle, for example a yearly pattern, that unfolds in an interday time dimension, there will be leap years with an additional day every fourth year. In such scenarios, we suggest to drop the additional observations at first from the leap year sub-periods and to continue as proposed in **M2**. Subsequently, for the leap years, if c represents the time point of the 29th of February of some leap year in the interday dimension, simply consider

$$\hat{m}(x_{1,c}, x_{2,b}) = \hat{m}(x_{1,c-1}, x_{2,b}). \tag{46}$$

Once the the seasonality surface $m(x_{1,a}, x_{2,b})$ has been estimated following either **M1** or **M2**, the residuals $\hat{r}_{a,b} = y_{a,b} - \hat{m}(x_{1,a}, x_{2,b})$ can be further decomposed as will be discussed in Section 5. In practice, to estimate the smoothed surfaces $\hat{m}$ we utilize the DCSmooth package in R for the DCS method, as discussed earlier in Section 3.1. The bandwidths for each of the N subperiods are manually selected with the most appropriate values for each year, ensuring satisfactory results for subsequent residual analysis.

5 Estimation method for intraday curves

In order to estimate the collection of intraday curves $(g_a(\cdot))$, $a = 1, \ldots, n_1$, we propose to use one-dimensional local polynomial smoothing similarly as discussed through the problems stated in Equations (12) to (17) to the observations whose seasonality surface has previously been removed through the estimation techniques described in Section 4. This series is denoted by $(\hat{r}_{a,b})$ in accordance with Model (2). Since the intraday curves are random and can behave quite differently even from day to day, we suggest to treat each day individually using an automated framework. Automated (one-dimensional) local polynomial regression under autocorrelated errors is well studied in Letmathe et al. (2024) under FARIMA errors and in Feng et al. (2020) generally under short-memory errors.

Since we allow long memory in our spatial residuals, the algorithm by Letmathe et al. (2024) could be preferable, however, in our particular scenario we do not find any evidence for long memory in the intraday dimension, as will be discussed in Section 7, which makes the algorithm by Feng et al. (2020) a suitable choice.

An option could be to estimate each intraday curve $g_a(\cdot)$ using a bandwidth $\hat{h}_a$ estimated particularly for day a. In such a case, each intraday series could potentially be smoothed using a different bandwidth. However, this approach is quite time-consuming, if there is data available for thousands of days. Moreover, we find that the estimated bandwidths usually do not differ strongly in our applications, which is why we propose instead to select a large enough number $K \in \mathbb{N}$ of days, $K \leq n_1$, for example $K = 100$, randomly (while setting a seed in standard statistical software for reproducibility). Given the set of randomly selected days $D \equiv \{a_1, a_2, \ldots, a_K\} \subseteq \{1, 2, \ldots, n_1\}$, estimate a bandwidth $\hat{h}_{a_i}$, $i = 1, \ldots, K$, for each of these days from the corresponding intraday curve. Then smooth each day $a = 1, \ldots, n_1$ using the average bandwidth $\hat{h}_{\text{ave}} = K^{-1} \sum_{i=1}^{K} \hat{h}_{a_i}$ to obtain the set of estimated intraday curves $(\hat{g}_a(\cdot))$.

Once $\hat{h}_{\text{ave}}$ has been identified, one may implement a fast matrix algorithm to obtain the set of estimated $g_a(\cdot)$, i.e. $(\hat{g}_a(\cdot))$ with $a = 1, \ldots, n_1$, in analogy to Definition 3.2 in (Schäfer, 2021b). Define $b_w = [n_2 \times \hat{h}_{\text{ave}} + 0.5]$ as the estimated absolute bandwidth, where $[\cdot]$ is the integer part, and

$$\mathbf{r}_{\text{left},a} = \begin{pmatrix} \hat{r}_{a,1} & \hat{r}_{a,2} & \ldots & \hat{r}_{a,2b_w+1} \end{pmatrix}^{\mathrm{T}} \tag{47}$$

is the $(2b_w + 1) \times 1$ vector containing the previously obtained $(\hat{r}_{a,b})$ needed for smoothing at a left boundary point on day a only. Then $\mathbf{R}_{\text{left}} = (\mathbf{r}_{\text{left},1} \vdots \mathbf{r}_{\text{left},2} \vdots \ldots \vdots \mathbf{r}_{\text{left},n_1})^{\mathrm{T}}$ is a $n_1 \times (2b_w + 1)$ matrix. Moreover, define $\mathbf{w}_b$ as the $(2b_w + 1) \times 1$ vector with realized weights for smoothing (following one-dimensional local polynomial regression) at time point b (unrequired weights, for example for smoothing at boundary points, are to be replaced by zero), i.e. it is the result of $\kappa_k! \iota_{\kappa_k+1}^{\mathrm{T}} \left(\mathbf{X}_k^{\mathrm{T}} \mathbf{W} \mathbf{X}_k\right)^{-1} \mathbf{X}_k^{\mathrm{T}}$ from Equation (21) in the analogous context of univariate local polynomial regression at a boundary point, and $\mathbf{W}_{\text{left}} = (\mathbf{w}_1 \vdots \mathbf{w}_2 \vdots \ldots \vdots \mathbf{w}_{b_w})$. Consequently,

$$\mathbf{G}_{\text{left}} = \mathbf{R}_{\text{left}} \mathbf{W}_{\text{left}} \tag{48}$$

is a simple matrix solution to obtain the smoothed curves at left boundary points. Analogous thoughts hold for the right boundary points, for which we consider $\mathbf{G}_{\text{right}}$ to contain the estimated intraday curves at the right boundary points. At interior points, define $\mathbf{w}_{\text{in}}$ as the $(2b_w + 1) \times 1$ weighting vector for an interior point in accordance with $\hat{b}_{\text{ave}}$. Let

$$\mathbf{r}_{a,b} = (\hat{r}_{a,b-b_w} \quad \hat{r}_{a,b-b_w+1} \quad \ldots \quad \hat{r}_{a,b+b_w})^{\mathrm{T}}, \quad b \in \{b_w + 1, b_w + 2, \ldots, n_2 - b_w\}, \tag{49}$$

be the vector with consecutive $\hat{r}_{a,b}$ on any fixed day a, where b is an interior point. Then $\mathbf{R}_b = (\mathbf{r}_{1,b} \vdots \mathbf{r}_{2,b} \vdots \cdots \vdots \mathbf{r}_{n_1,b})^{\mathrm{T}}$ and consequently

$$\mathbf{G}_b = \mathbf{R}_b \mathbf{w}_{\text{in}} \tag{50}$$

is a $n_1 \times 1$ vector containing the estimates of $g_a(x_{2,b})$ at time point b over all days a. Equation (50) is to be computed for all $b \in \{b_w + 1, b_w + 2, \ldots, n_2 - b_w\}$, separately. $\mathbf{G}_{\text{in}} = (\mathbf{G}_{b_w+1} \vdots \mathbf{G}_{b_w+2} \vdots \ldots \vdots \mathbf{G}_{n_2-b_w})$ then contains the estimated intraday curves for all interior points. Thus, $\mathbf{G} = (\mathbf{G}_{\text{left}} \vdots \mathbf{G}_{\text{in}} \vdots \mathbf{G}_{\text{right}})$ contains all elements of the estimated intraday curves. If $\hat{g}_a(x_{2,b}) = [\mathbf{G}]_{ab}$, where $[\mathbf{G}]_{ab}$ is the element in row a and column b of matrix $\mathbf{G}$, is the estimated value of the intraday curve at time b on day a, then $\hat{\tau}_{a,b} = \hat{r}_{a,b} - \hat{g}_a(x_{2,b})$ is the residual from the seasonality surface smoothing adjusted for the estimated intraday curve. $(\hat{\tau}_{a,b})$, $a = 1, \ldots, n_1$, $b = 1, \ldots, n_2$, is in our analyses assumed to be approximately stationary, so that they may be further studied through parametric approaches such as SFARIMA models.

We find that some of the intraday curves can behave erratically. As a countermeasure, we suggest to estimate the intraday curves using local cubic regression, which adapts better to drastic shifts in short amounts of time. A discussion of the existence of intraday curves and the empirical justification of the aforementioned approach is derived by empirical analysis of our example data in Section 7.

6 Imputation technique for spatial time series

The estimation method examined from the previous section presumes the availability of complete data. This should be a crucial factor for modeling and forecasting any time series. Nonetheless, in the field of environment in general (air temperature in particular and in our applied examples), where missing observations are common, there should be clarity about how the actual data are derived and how missing values are handled. To tackle this problem, there are several ways to deal with this situation as outlined in existing literature. This process is commonly referred to as "imputation". In practical terms, there exist different techniques for imputing missing values for one-dimensional, univariate time series, which is why, for the imputation step only, we treat our example series in Section 7 as collections of such series for the various observation days instead of spatial ones. It is known that climatic data are generally characterized by properties such as autocorrelation between time lags, seasonality, periodic cycles and the homogeneity effect over geographical areas. These characteristics could be useful information for developing imputation modelling schemes to fill the periods of missingness (Moritz et al., 2015; Moritz and Bartz-Beielstein, 2017; Yamoah et al., 2020). As outlined in the research introduced by Rubin (1976), missing data are commonly classified into three categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). It is important to highlight that missing values in air temperature in our scenarios typically adhere to the MAR mechanism. It is simply because that climate variables, specifically air temperature, are generally correlated with other variables, such as humidity, wind speed, cloud coverage levels, solar radiation, etc. and in instances where weather stations collect multiple climatic conditions, imputations for missing values in one missing variable can be inferred from the other observed variables. Most imputation methods assume MCAR and MAR because their missing data mechanisms are said to be ignorable, implying that the imputation modelling does not require special models for the cause of missingness and the likely values for the gaps (Rubin, 1976; Moritz et al., 2015). However, it is still sensible to estimate missing value information in our examples by employing at least one modeling technique. Thanks to the fact that the ratio of missing values in the applied examples are almost close to 0.03% in the whole time series, hence making a seemingly complicated imputation method unnecessary. For simplicity, the paper only focuses on considering one of the appropriate treatments of missingness. This method is selected from three common approaches: mean imputation, last observation carried forward (LOCF) and ARIMA model fitting combined with Kalman smoothing.

Let y_t be the observed (univariate, one-dimensional) time series with $t \in T^*$ with $T^* = \{t_1, t_2, \ldots, t_{n^*}\}$ being the set of available observation time points. Note that t_1 to t_{n^*} are all positive integers with $1 \leq t_1 < t_2 < \cdots < t_{n^*} \leq n$. By conclusion, the time points with actually available observations are not necessarily equidistant. On the contrary, the full set of observation time points, including those for which observations are missing, is denoted by $T = \{1, 2, \ldots, n\}$ so that $T^* \subseteq T$. Since $|T^*| = n^*$ and $|T| = n$, we conclude $m = n - n^*$ to be the number of missing observations. The first imputation approach considered, i.e. mean imputation, estimates the missing observations by

$$\hat{y}_t = (n^*)^{-1} \sum_{i=1}^{n^*} y_{t_i}, \quad \forall t \in \{T \backslash T^*\}, \tag{51}$$

while by the second approach, LOCF, we have

$$\hat{y}_t = y_{t-k}, \quad k = \min\left(\{x \in \mathbb{N}_+ : (t - x) \in T^*\}\right), \quad \forall t \in \{T \backslash T^*\}. \tag{52}$$

The most sophisticated time series imputation technique considered here is based on estimating univariate ARIMA models, i.e. an adaptation of Model (3), where $d_2 = p_2 = q_2 = 0$ with $n_2 = 1$ and $d_1 \in \mathbb{N}_0$, and thus reducing the model to a one-dimensional short-memory case. To account for missing observations, the estimation is conducted via QMLE following the model's state space representation. The imputations are subsequently obtained through Kalman smoothing (see e.g. Chapter 4 in Durbin and Koopman, 2012). Since spatial time series, also arranged as univariate time series, usually have huge numbers of observations and due to the fact that common estimation techniques such as QMLE exhibit enormous computation times under large numbers of observations, we propose to split the data y_t into R suitable smaller univariate time series $\left(y_t^{[i]}\right)$, $i = 1, \ldots, R$, of equal length n (also counting the missing observations) and with actual number of available observations n_i^*, correspondingly, where $t = (i-1)n+1, \ldots, in$, for example into time series for each day, and to apply the imputation technique to each subseries. Using the R package `imputeTS` (Moritz and Bartz-Beielstein, 2017) and the function `na_kalman()` with the option `model = "auto.arima"` the third imputation approach can be implemented.

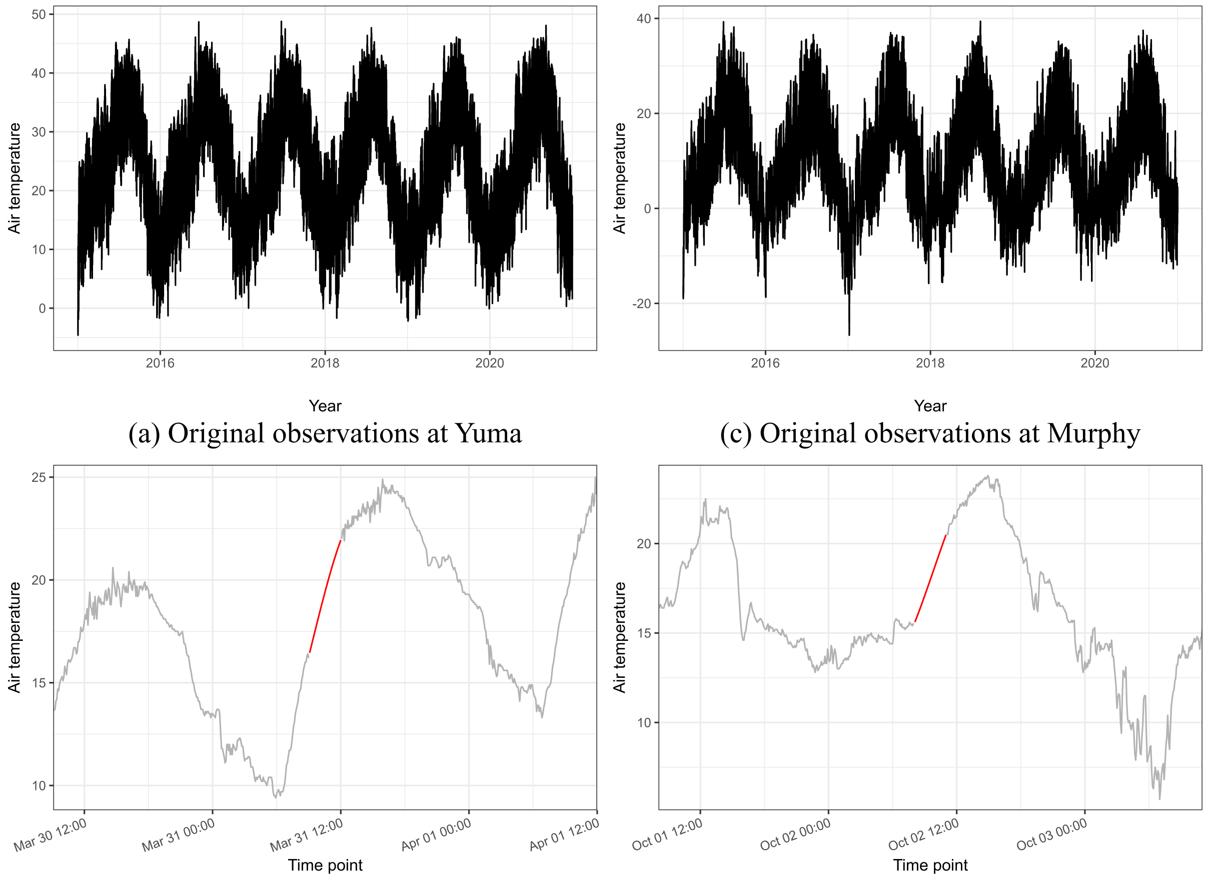
Upon evaluating various plots generated from these methods for the actual missing values in our example datasets, it was determined that employing the last mentioned ARIMA model fitting and Kalman smoothing method via `na_kalman()` (Moritz and Bartz-Beielstein, 2017) is the most appropriate approach. This conclusion is supported by several reasons: the missing values across the entire observation period are exceedingly minimal and their time points exhibit no systematic pattern. They are neither continuous over a substantial range of timestamps nor concentrated during typical times of the day across consecutive days. This decision is also supported by Moritz and Bartz-Beielstein (2017), who recommend using Kalman smoothing for imputation under more complex time series. In several instances, they demonstrated that the Kalman filter, however based on a general basic structural model instead of an explicit ARIMA form, yields more realistic imputation results (Moritz and Bartz-Beielstein, 2017) in comparison to other methods, for example using LOCF or mean imputation. Kalman smoothing also leads to reasonable results for our series as observable in Figure 1. Due to this fact and because the seemingly reasonable imputation results by `na_kalman()` for the relatively few missing observations are not expected to have a significant impact on the further Semi-SFARIMA estimation steps, we consider Kalman smoothing via `na_kalman()` throughout for imputation.

7 Application

7.1 Collection and preparation of the data

To illustrate the concept of identifying long-range dependence within observations on spatial surfaces and estimating them, environmental datasets are employed. This choice is primarily made because such datasets are typically accessible online at no cost. Since environmental data often exhibit missing values over extended periods with a large number of observations, ensuring the reliability of subsequent empirical findings is crucial. To achieve this, the selection of available data focuses on stations with minimal missing values. The dataset subjected to the proposed algorithm comprises two high-frequency environmental time series: air (average) temperature at the Yuma (Arizona) and Murphy (Idaho) station in the US. These variables are measured in Celsius degrees at 5-minute intervals from January 1, 2015, to December 31, 2020. The series are sourced directly from the National Oceanic and Atmospheric Administration (NOAA) website. This is an official website of the United State government, where environmental data, products, and services covering the depths of the ocean to the surface of the sun to drive resilience, prosperity, and equity for current and future generations are provided. The NCEI archive and backup copy contains more than 60 petabytes of data and is considered as one of the reliable websites for environmental data. In both datasets, there are consistent 5-minute observations throughout the day, totaling 288 observations daily, recorded from 00:00 to 23:55. Yuma's air temperature dataset comprises 631,296 observations, accounting for two leap years (2016 and 2020), which include the additional day February 29. This dataset has 231 missing values out of 631,296 observations, exhibits a minimal 0.04% missing ratio.

Likewise, Murphy's dataset on air temperature exhibits slightly fewer missing values, amounting to 206 instances, equating to a 0.03% missing rate. Both series are displayed



(a) Original observations at Yuma

(c) Original observations at Murphy

(b) Yuma - imputation on Mar 31, 2016

(d) Murphy - imputation on Oct 02, 2015

Figure 1: Illustration of imputation examples at Yuma and Murphy.

as univariate time series in Figures 1(a) and (c), respectively. Through this illustration, the yearly repeating pattern in air temperature is highlighted. Further details on the daily pattern cannot be derived from this representation, as those changes are too miniscule in order to be visible in Figures 1(a) and (c). Nonetheless, spatial representations of the (imputed) series are represented in Figures 3(a) and 4(a), from which the common pattern of air temperatures over the course of a day becomes more apparent.

Furthermore, Tables 1 and 2 summarize missing values at Yuma and Murphy stations, organized by month and year, respectively. Note that the largest sequence of missing values is in both series of length 36, which amounts to roughly three consecutive hours. To adress the missing observations, the imputation method described in Section 6 is applied to both datasets. Figures 1(b) and (d) illustrate the two selected examples of

Table 1: Missing values (NA) at Yuma station sorted by month-year.

Row	Subset	Month-Year	No.of NA
22286 - 22321	subset1	Mar - 2015	36
69301	subset2	Aug - 2015	1
131150 - 131185	subset3	Mar - 2016	36
210364 - 210366	subset4	Dec - 2016	2
225628 - 226453	subset5	Feb - 2017	3
228395 - 236017	subset6	Mar - 2017	56
271534	subset7	Jul - 2017	1
338510 - 338557	subset8	Mar-2018	48
445646 - 445693	subset9	Mar-2019	48

Table 2: Missing values (NA) at Murphy station sorted by month-year.

Row	Subset	Month-Year	No.of NA
10117	subset 1	Feb - 2015	1
79010 - 79045	subset 2	Oct - 2015	36
96370	subset 3	Dec - 2015	1
107005 - 107014	subset 4	Jan - 2016	5
117049	subset 5	Feb - 2016	1
176150 - 176185	subset 6	Sep - 2016	36
282662 - 282697	subset 7	Sep - 2017	36
369248	subset 8	Jul- 2018	1
381230 - 384097	subset 9	Aug - 2018	37
467570 - 467617	subset 10	Jun - 2019	48
492391	subset 11	Sep - 2019	1
591324	subset 12	Aug - 2020	1
606589	subset 13	Oct - 2020	1
614577	subset 14	Nov - 2020	1

the imputation. In Figure 1(b), the imputation in the Yuma series on March 31, 2026, is displayed, whereas in Figure 1(d), the imputation results on October 02, 2026, are shown for the Murphy series. In both cases, 36 consecutive missing values were being replaced. Figures 1(b) and (d) support that the considered imputation works as intended: the red lines connect the observed parts of the series in a reasonable way and also exhibit a certain level of smoothness. Consequently, we continue forward under consideration of the obtained imputed series.

7.2 Model estimation

7.2.1 Air temperature at Yuma station

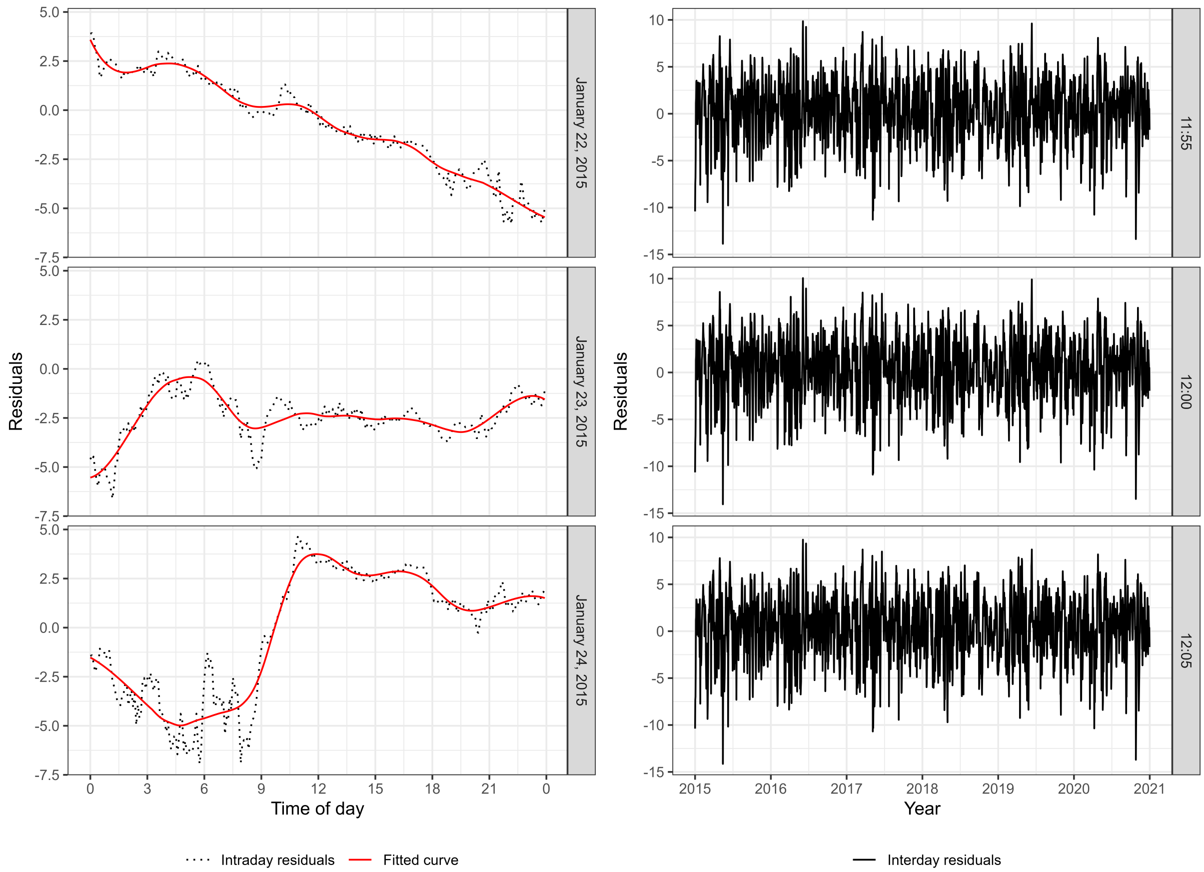
For the Yuma dataset, we estimate the Semi-SFARIMA following an initial estimation of the spatial seasonality following method **M1** as presented in Section 4, i.e. a different set of bandwidths is employed for each year of data. The manually chosen bandwidths according to this method for smoothing the temperature surface at Yuma station is summarized in the Table 3. For h_1, for the interday dimension, the bandwidths all lie on the interval

Table 3: The chosen bandwidths for smoothing temperature surface at Yuma.

Year	h_1	h_2
2015	0.1124	0.1451
2016	0.1124	0.1272
2017	0.1231	0.1415
2018	0.1130	0.1377
2019	0.1157	0.1448
2020	0.1083	0.1325

from 0.0854 to 0.1039, while for h_2, for the intraday dimension, they lie between 0.0883 and 0.0967. Hence, the bandwidths do not differ noticeably between the yearly subsets. As previously noted regarding the cyclical patterns, since there is no identified trend stationarity in this context, applying a smoothing surface across the entire period (i.e., six years) is not feasible. This leads to the persistence of cyclical structures in the residuals.

On the contrary, when employing DCS for each year independently, the seasonal cycles



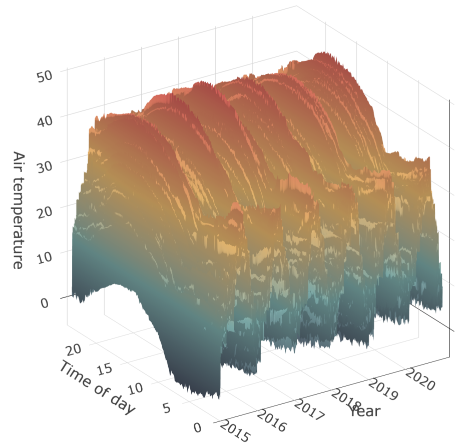
(a) Examples of intraday residuals

(b) Examples of interday residuals

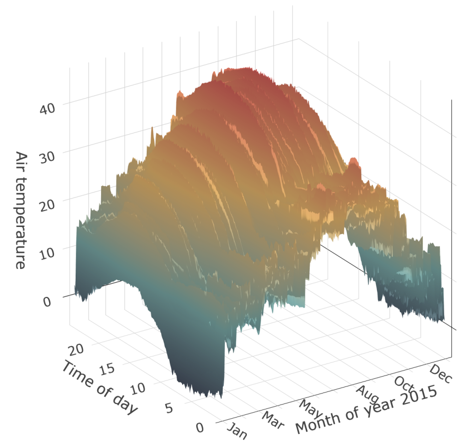
Figure 2: An illustration of selected consecutive intraday and interday residual series. The intraday residuals are shown alongside individually fitted curves.

identified previously do not persist in the residuals, resulting in a greater degree of stability across the entire series, as shown in Figure 3(d). As an example, the estimated seasonal surface for the year 2015 is displayed in Figure 3(c). It captures the expected patterns, i.e. a drop in temperature in the early morning hours, followed by an increase in temperature until noon or the early afternoon, and then ending in a reduction in temperature for the rest of the day for the intraday dimension, and a change in temperature in accordance with the natural seasons in the interday dimension. As observed in Figure 3(d), the seasonal patterns atop each year are absent in the residuals.

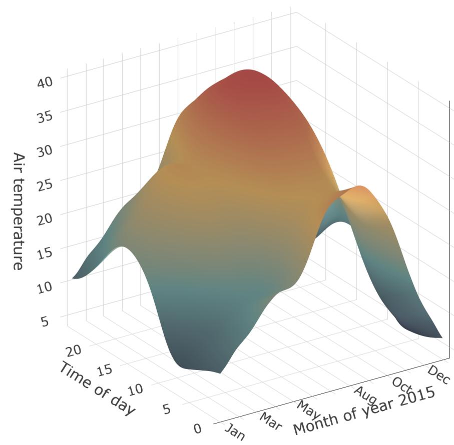
However, as can be deduced from Figure 3(a), the obtained residuals cannot be considered to be roughly stationary. While slices of interday residuals as shown in Figure 2(b) appear to be roughly stationary, sliced intraday residuals admit nonstationary behavior



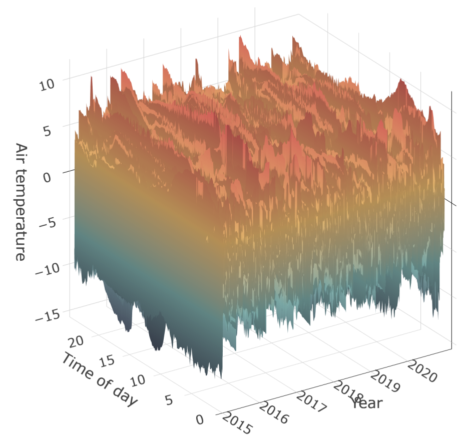
(a) Original observations (2015-2020)



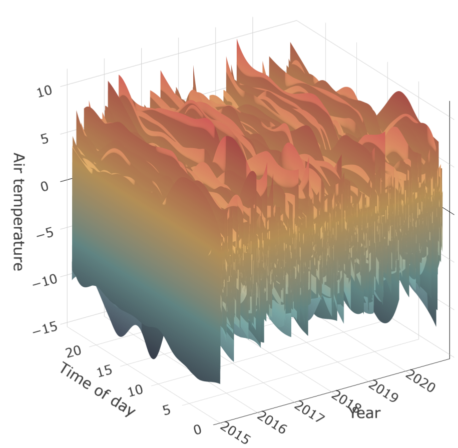
(b) Original observations (2015)



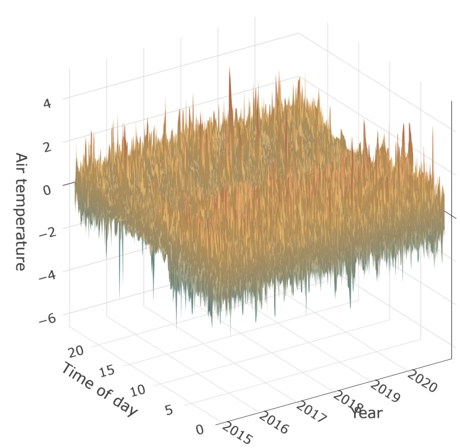
(c) Smoothed surface (2015)



(d) Residuals (2015-2020)



(e) Intraday curves (2015-2020)



(f) Adjusted residuals (2015-2020)

Figure 3: Some figures illustrating empirical results at Yuma station (2015-2020).

that also varies strongly from day to day as shown in Figure 2(a). This behavior warrants the estimation of individual intraday curves as described in Section 5.

For this purpose, we consider the algorithm by Feng et al. (2020) under consideration of local cubic regression and apply it to 100 randomly selected intraday series of the Yuma spatial series, from which the estimated seasonality surface has been removed beforehand. The resulting estimated average bandwidth is $\hat{h}_{\text{ave}} = 0.1357$, which is subsequently used in local cubic regressions for all of the intraday series. For the selected examples of intraday residuals in Figure 2(a), the corresponding estimated curves are shown as well. They seem to capture the intraday movements suitably. In Figure 3(e), the collection of estimated intraday curves is displayed, whereas the spatial series adjusted for intraday curves is illustrated in Figure 3(f). A direct comparison between Figures 3(d) and 3(f) shows that the newly obtained residuals appear to be more stable than the observations solely adjusted for seasonality. Nonetheless, Figure 3(f) shows that the level of variation in the residuals is stronger during the night than during the day throughout. Notwithstanding that this phenomenon could be modelled through approaches that account for heteroskedasticity, further discussion on this topic is omitted here.

Now that the residuals in Figure 3(f) are available, we can fit a spatial FARIMA to them. The results of the estimated parameters for the spatial FARIMA$(1,1)(1,1)$ for Yuma's dataset is summarized in Table 4. Firstly, it can be observed that $\hat{d}_2 = 0$, i.e. we are unable to find long memory in the intraday dimension, while $\hat{d}_1 = 0.1584$, which indicates mild long-range dependence in the interday dimension. Secondly, the estimates $\hat{\phi}_1 = 0.4748$ and $\hat{\phi}_2 = 0.7029$ show a positive AR-coefficient in each dimension, with the coefficient being clearly larger for the intraday dimension. Thirdly, the estimated MA-parameter in the interday dimension $\hat{\theta}_1 = -0.6087$ is moderately negative, whereas $\hat{\theta}_2 = 0.0413$ is weakly positive. Additionally, the (asymptotically valid) standard errors of the estimated parameters are calculated using the formula proposed by Beran et al. (2009), which are also shown in Table 4. The standard errors are all quite small, indicating statistical significance at common significance levels for all parameters except $\hat{d}_2$ due to $\hat{d}_2 = 0$. Moreover, the estimated effect of θ_2 is quite close to 0 in magnitude and therefore possibly negligible.

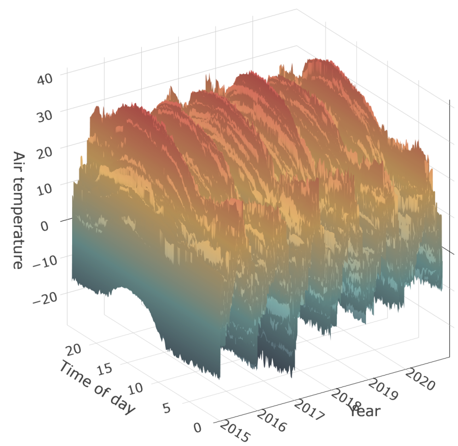
Table 4: The estimated SFARIMA(1, 1)(1, 1) parameters for the Yuma residuals.

Parameter	ϕ_1	ϕ_2	d_1	d_2	θ_1	θ_2	σ
Estimated value	0.4748	0.7029	0.1584	0	-0.6087	0.0413	0.3077
Standard error	0.0024	0.0013	0.0013	0.0016	0.0026	0.0008	

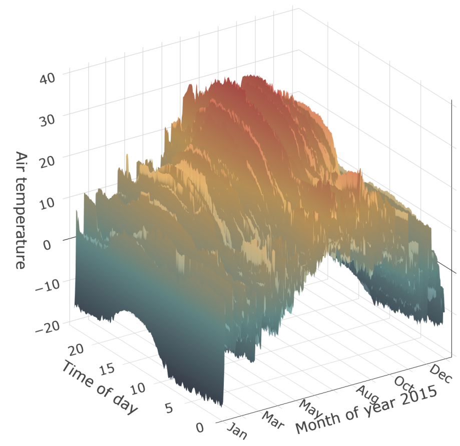
7.2.2 Air temperature at Murphy station

In contrast, for analyzing Murphy's dataset the estimation method **M2** is utilized. According to the methodology outlined in Section 4, a breakdown is as follows: firstly, the smoothing surface for each year is executed in the usual manner, with consistent bandwidths chosen for every year ($h_1 = 0.0725$ and $h_2 = 0.1075$). Subsequently, the mean smoothed surface for the six-year period is computed by amalgamating these smoothed matrices into a comprehensive six-year matrix, excluding the smoothed values for February 29 in the years 2016 and 2020. The residuals for each year are then manually derived by deducting the mean smoothed surface (obtained in the preceding step) from the original observations. For the leap years 2016 and 2020, the values of the average smoothed surface for February 28 are duplicated for February 29 to maintain consistency with the number of observation in the original dataset. The reasoning behind this decision is that it is expected that the values of the surface do not change drastically within a day and should be roughly identical between February 28 and February 29.

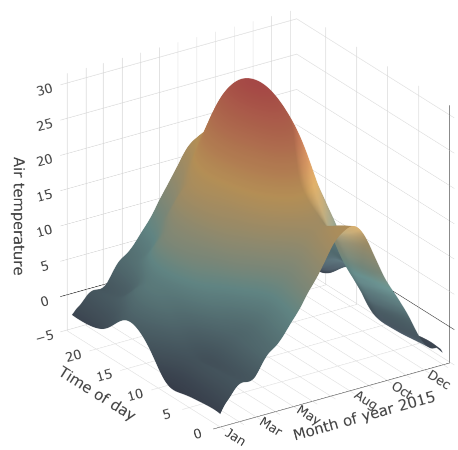
The average surface (for a year with 365 days) is illustrated in Figure 4(c). We can observe similar patterns as for the Yuma surface of year 2015 discussed in the previous subsection. Subsequently, the residual plot resulting from this proposed procedure is depicted under Figure 4(d). Analogously to the residuals obtained for the Yuma series, the residuals for the Murphy series still show notabole peaks and drops caused by random intraday effects, while the interday residuals are approximately stationary. A plot of selected intraday and interday residual series for the Murphy data is omitted here. As a consequence, an estimation of intraday curves is employed with (one-dimensional) local cubic regression, where a random selection of 100 intraday series results in $\hat{h}_{\text{ave}} = 0.1334$ following the average of the bandwidths selected by Feng et al. (2020). The collection of estimated intraday curves is shown in Figure 4(e), while the corresponding residuals are being depicted in Figure 4(f). The residuals are now approximately stationary and



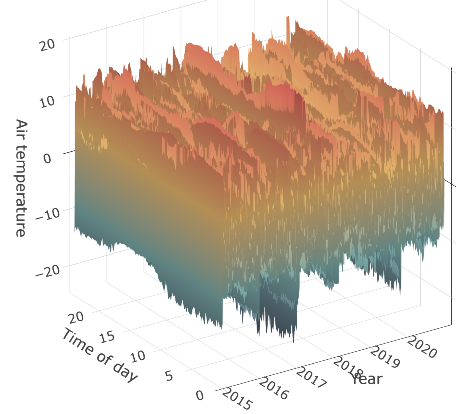
(a) Original observations (2015-2020)



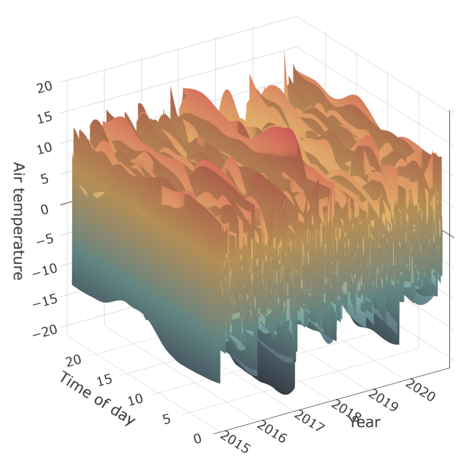
(b) Original observations (2015)



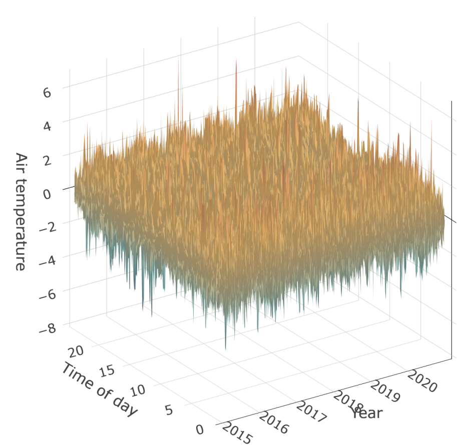
(c) Smoothed surface (average)



(d) Residuals (2015-2020)



(e) Intraday curves (2015-2020)



(f) Adjusted residuals (2015-2020)

Figure 4: Some figures illustrating empirical results at Murphy station (2015-2020).

Table 5: The estimated SFARIMA$(1,1)(1,1)$ parameters for the Murphy residuals.

Parameter	ϕ_1	ϕ_2	d_1	d_2	θ_1	θ_2	σ
Estimated value	0.4542	0.6544	0.0960	0	-0.5344	0.1247	0.4666
Standard error	0.0045	0.0013	0.0012	0.0014	0.0048	0.0007	

do, in comparison to the Yuma residuals in Figure 3(f), not show any visible signs of heteroskedasticity.

Lastly, the ultimate phase entails modeling the residuals using a spatial FARIMA$(1,1)$ $(1,1)$ model. The estimated parameters, again, are summarized in Table 5 below. We have $\hat{d}_1 = 0.0960$ and $\hat{d}_2 = 0$, which indicates once again an absence of long memory in the intraday dimension and a mild long-range dependence in the interday dimension. Yet again, the two AR parameters $\hat{\phi}_1 = 0.4542$ and $\hat{\phi}_2 = 0.6544$ are positive, with the intraday AR-parameter being larger than the one for the interday dimension. $\hat{\theta}_1$ is moderately negative and $\hat{\theta}_2 = 0.1247$ weakly positive. The obtained (asymptotically valid) standard errors also visible in Table 5 lead to the conclusion that all estimated parameters are statistically significant with the exception of $\hat{d}_2$.

7.3 Findings

The proposed methods **M1** and **M2** result in suitable estimated seasonal surfaces as shown in context of the Yuma and Murphy time series. They manage to acertain that the dual-seasonal pattern in the combined interday and intraday dimensions is captured well enough in analogy to expectations regarding the course of temperature over the course of a day and of a year. The proposed method to identify the random intraday curves present in the data works well on the surface-adjusted data, so that the resulting residual series are approximately stationary and therefore suitable for further analysis via parametric spatial time series models. While both methods M1 and M2 give in fact similar results, method M1 should generally be preferred, as the idea of a change in seasonal pattern over the years is more realistic than a fixed, yearly repeating spatial seasonal pattern. If, however, an exactly periodic pattern is to be expected, method M2 will surely provide more reliable results. In the given scenarios, both approaches resulted in suitable estimates.

On a further note, we are generally unable to find long memory in the intraday dimension of the analyzed temperature series, whereas there is in both cases mild long memory in the interday dimension. Under consideration of (41), further hypothesis tests on the equality of corresponding parameters of the two dimensions, i.e. of $d_1 = d_2$, $\phi_1 = \phi_2$ and $\theta_1 = \theta_2$, lead for both example series in each case to a p-value of approximately zero, indicating that a spatial model could be more preferable than a simple univariate model due to differences in dependency structure for both dimensions. That aside, generally, even though two different estimation methods were used for the Yuma and Murphy data regarding the seasonal surface, we find strong similarities in the fitted SFARIMA$(1,1)(1,1)$ models for both residual series. The AR-, MA-, and long-memory parameters are all quite similar, even almost identical, between the two models, meaning that the air temperatures in Yuma and Murphy follow similar patterns.

8 Concluding remarks

The discussed methods provide a first suggestion for handling two-dimensional seasonal patterns and random intraday curves in spatial time series. The two methods for seasonal patterns are (i) a year-wise smoothing of the surface using LP-DCS with varying sets of bandwidths, where the resulting surfaces are considered for the respective years, and (ii) a year-wise smoothing of the surface using LP-DCS with the same set of bandwidths, from which an average surface is obtained that is the estimated seasonal surface for each year. Thereafter, we used one-dimensional local cubic smoothing to identify individual intraday curves. Subsequently, we fitted spatial long-memory time series models to the (spatial) residual series resulting in clear signs of differences in the marginal models and of the existence of long memory in at least some of the dimensions, making the SFARIMA a favorable choice over common one-dimensional short-memory time series models.

Nonetheless, some aspects are left open for future research: since the observed example series had missing values, an imputation technique for one-dimensional time series was employed and it was assumed that, due to the small number of missing values, the effect of the imputation was negligible. The development of imputation methods for spatial time series, preferably under some long memory design, could lead to a theoretically more justified approach. In addition, the characteristics of the proposed estimation methods

M1 and M2, which make use of the LP-DCS, need to be evaluated. Linked to this is that we surmised that the asymptotic results of the LSE for a SFARIMA also hold approximately, when considered not directly for observations but for residuals after a preliminary surface (and intraday curves) estimation. While there are indications for this to be true, actual proof of this is currently missing in existing literature. Moreover, we found existing bandwidth selection algorithms for LP-DCS under spatial time series with long memory to work worse than expected, assumingly due to the random intraday curves, which led us to select bandwidths manually throughout. Further investigation of such algorithms and possible improvements of them are therefore a necessity, which could generate even more reliable empirical results. Further important research should cover topics such as forecasting, i.e. point and interval forecasts (both under normal and non-normal innovations), under a Semi-SFARIMA.

References

Andel, J. (1986). Long Memory Time Series models. In: *Kybernetika (Prague)* 22.2, pp. 105–123.

Andersen, T. and Bollerslev, T. (1997). Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns. In: *The Journal of Finance* 52.3, pp. 975–1005.

Baillie, R. and Chung, H. (2002). Estimation of GARCH Models from the Autocorrelations of the Squares of a Process. In: *Journal of Time Series Analysis* 22.6, pp. 631–650.

Beran, J. (1994). *Statistics for Long-Memory Processes.* 1st ed. Chapman & Hall CRC.

Beran, J. and Feng, Y. (2002a). Iterative plug-in algorithms for SEMIFAR models Definition, convergence, and asymptotic properties. In: *Journal of Computational and Graphical Statistics* 11.3, pp. 690–713.

– (2002b). Local polynomial fitting with long-memory, short-memory and antipersistent errors. In: *Annals of the Institute of Statistical Mathematics* 54.2, pp. 291–311.

Beran, J., Feng, Y. and Hebbel, H. (2015). *Emperical Economic and Financial Research: Theory, Methods and Practice.* Cham, Switzerland: Springer.

Beran, J., Ghosh, S. and Schell, D. (2009). On least squares estimation for long-memory lattice processes. In: *Journal of Multivariate Analysis* 100.10, pp. 2178–2194.

Ding, Z., Granger, C. and Engle, R. (1993). A long memory property of stock market returns and a new model. In: *Journal of Empirical Finance* 1, pp. 83–106.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Models.* 2nd ed. Oxford, United Kingdom: Oxford University Press.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. In: *The Annals of Statistics* 20.4, pp. 2008–2036.

– (1996). Local polynomial modelling and its applications. In: *Monographs on Statistics and Applied Probability* 66.

Feng, Y., Beran, J., Ghosh, S., Letmathe, S. and Schäfer, B. (2021). *A semiparametric generalization of spatial fractional ARIMA processes.* Tech. rep. unpublished. Paderborn University.

Feng, Y. and Peitz, C. (2013). *Double-conditional smoothing of high-frequency volatility surface in a semiparametric spatial multiplicative component GARCH.* Tech. rep. Unpublished. Paderborn University.

Feng, Y., Gries, T. and Fritz, M. (2020). Data-driven local polynomial for the trend and its derivatives in economic time series. In: *Journal of Nonparametric Statistics* 32.2, pp. 510–533.

Gasser, T. and Müller, H. (1979). Kernel Estimation of Regression Functions. In: *Smoothing Techniques for Curve Estimation. Lecture Notes in Mathematics 757*. Ed. by T. Gasser and M. Rosenblatt. 1st ed. Berlin, Germany: Springer, pp. 23–68.

Giraitis, L. and Leipus, R. (1995). A generalized fractionally differencing approach in long-memory modeling. In: *Lithuanian Mathematical Journal* 35.1, pp. 53–65.

Granger, C. and Ding, Z. (1996). Varieties of long memory models. In: *Journal of Econometrics* 73.1, pp. 61–77.

Granger, C. and Joyeux, R. (1980). An Introduction to Long Memory Time Series models and Fractional Differencing. In: *Journal of Time Series Analysis* 1.1, pp. 15–29.

Gray, H., Zhang, L. and Woodward, W. (1989). On Generalized fractional Processes. In: *Journal of Time Series Analysis* 10.3, pp. 233–257.

– (1994). On Generalized fractional Processes - A correction. In: *Journal of Time Series Analysis* 15.5, pp. 561–562.

Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. 1st ed. Cambridge, UK: Cambridge University Press.

Hallin, M., Lu, Z. and Tran, L. (2004). Local linear spatial regression. In: *The Annals of Statistics* 32.6, pp. 2469–2500.

Härdle, W. and Müller, M. (2000). Multivariate and Semiparametric Kernel Regression. In: *Smoothing and Regression: Approaches, Computation, and Application*. Ed. by M. G. Schimek. 1st ed. New York, NY: John Wiley & Sons, pp. 357–391. DOI: 10.1002/9781118150658.ch12.

Hassler, U. (1994). (Mis)specification of long memory in seasonal time series. In: *Journal of Time Series Analysis* 15.1, pp. 19–30.

Herrmann, E., Engle, J., Gasser, T. and Wand, M. (1995). A Bandwidth Selector for Bivariate Kernel Regression. In: *Journal of the Royal Statistical Society* 57, pp. 171–180.

Hosking, M. (1981). Fractional differencing. In: *Biometrika* 68.1, pp. 165–176.

Hubbard, K. and You, J. (2005). Sensitivity analysis of quality assurance using the spatial regression approach - A case study of the maximum/minimum air temperature. In: *Journal of Atmospheric and Oceanic Technology* 22.10, pp. 1520–1530.

Letmathe, S., Beran, J. and Feng, Y. (2024). An extended exponential SEMIFAR model with application in R. In: *Communications in Statistics - Theory and Methods* 52.22, pp. 7914–7926.

Lin, Z., Mo, X. and Li, H. (2002). Comparison of three spatial interpolation methods for climate variables in China. In: *Acta Geographica Sinica* 57.1, pp. 47–56.

Martin, R. (1979). A subclass of lattice processes applied to a problem in planar sampling. In: *Biometrika* 66.2, pp. 209–217.

Moritz, S. and Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. In: *The R Journal* 9.1, pp. 207–218. URL: `https://cran.r-project.org/package=imputeTS`.

Moritz, S., Sarda, A., Bartz-Beielstein, T., Zaefferer, M. and Stork, J. (2015). Comparison of Different Methods for Univariate Time Series Imputation in R. Preprint published on arXiv. DOI: `10.48550/arXiv.1510.03924`.

Müller, H. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. New York, NY: Springer.

– (1992). Change-points in nonparametric regression analysis. In: *The Annals of Statistics* 20.2, pp. 737–761.

Müller, H. and Prewitt, K. (1993). Multiparameter Bandwidth Processes and Adaptive Surface Smoothing. In: *Journal of Multivariate Analysis* 47, pp. 1–21.

Müller, H. and Wang, J. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. In: *Biometrics* 50, pp. 61–76.

Peitz, C. and Feng, Y. (2015). Double Conditional Smoothing of High-Frequency Volatility Surface Under a Spatial Model. In: *Empirical Economic and Financial Research: Theory, Methods and Practice. Advanced Studies in Theoretical and Applied Econometrics 48*. Ed. by J. Beran, Y. Feng and H. Hebbel. 1st ed. Cham, Switzerland: Springer, pp. 341–356.

Porter-Hudak, S. (1990). An Application of the Seasonal Fractionally Differenced Model to the Monetary Aggregates. In: *Journal of the American Statistical Association* 85.410, pp. 338–344.

Reynolds, R. and Smith, T. (1994). Improved global sea surface temperature analyses using optimum interpolation. In: *Journal of Climate* 7.6, pp. 929–948.

Rubin, D. (1976). Inference and Missing Data. In: *Biometrika* 63.3, pp. 581–592.

Ruppert, D. and Wand, M. (1994). Multivariate Locally Weighted Least Squares Regression. In: *The Annals of Statistics* 22.3, pp. 1346–1370.

Schäfer, B. (2021a). *DCSmooth: Nonparametric Regression and Bandwidth Selection for Spatial Models*. R package version 1.0.2. Package archived on CRAN. URL: `https://cran.r-project.org/package=DCSmooth`.

– (2021b). *Non- and Semiparametric Methods for Surface Estimation of Functional Time Series on a Lattice*. Doctor Dissertation. Paderborn University.

Scott, D. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken, NJ: John Wiley & Sons.

Snell, S., Gopal, S. and Kaufmann, R. (2000). Spatial interpolation of surface air temperatures using artificial neural networks: Evaluating their use for downscaling GCMs. In: *Journal of Climate* 13.5, pp. 886–895.

Stahl, K., Moore, R., Floyer, J., Asplin, M. and McKendry, I. (2006). Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. In: *Agricultural and Forest Meteorology* 139.3–4, pp. 224–236.

Wang, H. and Wang, J. (2009). Estimation of the trend function for spatio-temporal models. In: *Journal of Nonparametric Statistics* 21.5, pp. 567–588.

Wang, J., Christakos, F. and Hu, M. (2009). Modeling spatial means of surfaces with stratified nonhomogeneity. In: *IEEE: Transactions on Geoscience and Remote Sensing* 47.12, pp. 4167–4174.

Woodward, W., Cheng, Q. and Gray, H. (1998). A k-factor GARMA long-memory model. In: *Journal of Time Series Analysis* 19.4, pp. 485–504.

Yamoah, E., Müller, U., Taylor, S. and Fisher, A. (2020). Missing data imputation of high-resolution temporal climate time series data. In: *Meteorological Applications* 27.1, pp. 1–18.

Yang, L. and Tschernig, R. (1999). Multivariate bandwidth selection for local linear regression. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 61, pp. 793–815.

You, J., Hubbard, K. and Goddard, S. (2008). Comparison of methods for spatially estimating station temperatures in a quality control system. In: *International Journal of Climatology* 28.6, pp. 777–787.

Recent discussion papers

2026-03	Dominik Schulz Thi Thu Huong Do Yuanhua Feng	A semiparametric spatial FARIMA applied in the presence of spatial seasonality
2026-02	Dominik Schulz	The R Package deseats for Data-Driven Trend and Seasonality Estimation in Time Series
2026-01	Dominik Schulz Yuanhua Feng Thomas Gries Marlon Fritz Sebastian Letmathe	Diagnosing the trend and bootstrapping the forecasting intervals using a semiparametric ARMA
2025-06	Sarah Kühn Nadja Stroh-Maraun	Community Costs in Neighborhood Help Problems
2025-05	Sarah Kühn	Child Care Allocation Mechanisms: Navigating Incomplete Preference Elicitation
2025-04	Sarah Kühn Papatya Duman Britta Hoyer Thomas Streck Nadja Stroh-Maraun	Non-induced Preferences in Matching Experiments
2025-03	Marlon Fritz Thomas Gries Lukas Wiechers	Growth with Mismatch: Theory and Evidence from TFP Estimates
2025-02	Lukas Wiechers	A Real-Time Analysis of Fundamentals and Bubbles in the S&P 500
2025-01	Claus-Jochen Haake Martin R. Schneider	An Axiomatization of the Banzhaf Index to Measure Influence in Qualitative Comparative Analysis
2024-05	Iuliia Myroshnychenko Thomas Gries	Application of a dynamic welfare-centered energy transition model
2024-04	Claus-Jochen Haake Thomas Streck	A Measure for Contestedness of a Two-Person Bargaining Problem
2024-03	Youssouf Sangare Edem Adotey	Institutions and challenges to development: An Analysis of Economic Growth, Governance, and Development in Sub-Saharan Africa
2024-02	Papatya Duman Claus-Jochen Haake	Size Reduction Reform in German Parliament: a game theoretic analysis of power indices in the Bundestag
2024-01	Iuliia Myroshnychenko Thomas Gries	Developing a dynamic welfare-centered energy transition model: methodological challenges and perspectives
2023-03	Yuanhua Feng Thomas Gries Sebastian Letmathe	FIEGARCH, modulus asymmetric FILog-GARCH and trend-stationary dual long memory time series
2023-02	Claus-Jochen Haake Thomas Streck	Distortion through modeling asymmetric bargaining power
2023-01	Carina Burs	A Model of Cycles and Bubbles under Heterogeneous Beliefs in Financial Markets
2022-06	Marlon Fritz Thomas Gries Lukas Wiechers	An Early Indicator for Anomalous Stock Market Performance
2022-05	Sebastian Letmathe	Data-driven P-Splines under short-range dependence