Available online at www.sciencedirect.com
ScienceDirect
Mathematics and Computers in Simulation 146 (2018) 27–43
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Original articles

Robust estimation of fractional seasonal processes: Modeling and forecasting daily average SO_2 concentrations

Valdério Anselmo Reisen[a,b,d,*], Edson Zambon Monte[b], Glaura da Conceição Franco[c], Adriano Marcio Sgrancio[b], Fábio Alexander Fajardo Molinares[a], Pascal Bondon[d], Flávio Augusto Ziegelmann[e], Bovas Abraham[f]

[a] *Department of Statistics, Federal University of Espírito Santo, Espírito Santo, Brazil*
[b] *Graduate Program in Environmental Engineering, Federal University of Espírito Santo, Espírito Santo, Brazil*
[c] *Department of Statistics, Federal University of Minas Gerais, Minas Gerais, Brazil*
[d] *Laboratoire des Signaux et Systèmes, CNRS, Université Paris-Saclay, France*
[e] *Department of Statistics, Federal University of Rio Grande do Sul, Rio Grande do Sul, Brazil*
[f] *Department of Statistics, University of Waterloo, Ontario, Canada*

Received 13 January 2016; received in revised form 4 August 2017; accepted 8 October 2017
Available online 7 November 2017

Abstract

This paper deals with the estimation of seasonal long-memory time series models in the presence of 'outliers'. It is long known that the presence of outliers can lead to undesirable effects on the statistical estimation methods, for example, substantially impacting the sample autocorrelations. Thus, the aim of this work is to propose a semiparametric robust estimator for the fractional parameters in the seasonal autoregressive fractionally integrated moving average (SARFIMA) model, through the use of a robust periodogram at both very low and seasonal frequencies. The model and some theories related to the estimation method are discussed. It is shown by simulations that the robust methodology behaves like the classical one to estimate the long-memory parameters if there are no outliers (no contamination). On the other hand, in the contaminated scenario (presence of outliers), the standard methodology leads to misleading results while the proposed method is unaffected. The methodology is applied to model and forecast sulfur dioxide (SO_2) pollutant concentrations which have seasonal long-memory features and occasional large peak pollutant concentrations.

Keywords: Forecasting; Robust periodogram; Outliers; SO_2 pollutant; Long-memory

* Correspondence to: Department of Statistics, Federal University of Espírito Santo, 29075-910, 514, Vitoria, ES, Brazil.
E-mail address: valderio.reisen@ufes.br (V.A. Reisen).

https://doi.org/10.1016/j.matcom.2017.10.004

1. Introduction

Sulfur dioxide (SO_2) is part of a group of highly reactive gases known as "oxides of sulfur". The main sources of SO_2 emissions into the atmosphere come from fossil fuels (coal and petroleum products), power and other industrial plants. Other emission sources of SO_2 are as follows: industrial processes such as steel and mining, burning fuels containing high sulfur level due to transportation vehicles such as locomotives and large ships, pulp industry, and natural sources such as volcanic emissions (Andersson et al. [1]). High levels of pollution emissions can result in peaks of SO_2 concentration in the atmosphere.

The high SO_2 concentrations are associated with several effects on population health, as for instance the increase of chronic bronchitis (Kanaroglou et al. [23]). The exposure to SO_2, even for short periods of time (ranging from 5 min to 24 h), may cause respiratory problems, including bronchoconstrictor and increased asthma symptoms. Consequently, these can cause an increase in hospital admissions of the population at high risk, such as elderly, children and asthmatics.

Several statistical methods have been employed to study the impacts of SO_2 pollution on the environment, such as: principal component and cluster analysis (Cheng and Lam [10]); artificial neural networks for predicting sulfur dioxide concentrations (Saral and Ertürk [48]) and other pollutants (Brunelli et al. [6]); multiple regression models to investigate the influence of emission sources and meteorological conditions on SO_2 pollution (Luvsan et al. [28]); among others. In terms of forecasting, statistical models based on multiple regression and time series tools, such as the autoregressive integrated moving average (ARIMA) model have been widely used.

Many time series data exhibit a seasonality pattern. For pollutants, seasonal variation is often associated with the changes in meteorological parameters. Furthermore, the reduction of emission levels during weekends can also induce seasonality in the series. Thus, it is very important to consider statistical tools which take into account the seasonality effect.

More recently, several authors have studied time series with long-range dependence (often named long memory property). In time domain analysis, the long-memory dependence is generally characterized by a slow decay of the autocorrelation function. Granger and Joyeux [17] and Hosking [19] proposed the autoregressive fractionally integrated moving average (ARFIMA) model to describe the long-memory feature of a time series. The ARFIMA process is an extension of the ARIMA, where the parameter of integration d assumes fractional values. The methods proposed for estimating the parameters of the ARFIMA model are classified either as parametric or semiparametric. Parametric methods consist of simultaneous estimation of model parameters, usually by maximum likelihood. On the other hand, semiparametric estimation is performed in two steps: firstly the parameter d is estimated, and, in a second step, using the estimated d from the first step, the autoregressive and moving average parameters are estimated. The most popular semiparametric estimator was proposed in Geweke and Porter-Hudak [15]. For a recent review on this subject, see Palma [33], for instance. Modifications of this estimator have been developed by Reisen [39], Lobato and Robinson [27], Delgado and Robinson [11], Velasco [54], among others. More specifically, Sena et al. [50], Molinares et al. [31] and Reisen et al. [40] investigated the estimator under various model specifications, such as the presence of non-gaussian errors and outliers.

Due to the important features of the ARFIMA model, it has often been used to analyze environmental problems. As an example, Iglesias et al. [22] adopted an ARFIMA model as the underlying generating process of pollutant concentrations which present long memory and missing values. These phenomena are widely found in time series of different areas of interest.

An extension of the ARFIMA to handle time series with seasonality is the SARFIMA model, which has already been addressed in the literature by several authors. Porter-Hudak [35] applied the seasonal fractionally differenced model to study the monetary aggregates of the United States. Reisen et al. [44] estimated the fractional and seasonal parameters of SARFIMA models by means of a semiparametric procedure, considering a nonconstant conditional error variance. Reisen et al. [45] studied the properties of the SARFIMA model when the data exhibits one and two seasonal periods and short-memory components. Ye et al. [55] proposed a new method to estimate the fractional difference parameter in the SARFIMA model using tapered periodogram. Hsu and Tsai [20] proposed a semiparametric estimation method for seasonal long-memory time series, by considering a generalized exponential model in the frequency domain. Chan and Tsai [9] have studied the autocorrelation structure and the spectral density function of SARFIMA processes in aggregated discrete-time process, using the Whittle estimator. Tsai at al. [53] evaluated the properties of the Whittle likelihood estimation in SARFIMA models in the presence of measurement errors.

Furthermore, environmental time series often contain unusual observations influenced by external events that can cause changes in their dynamics. These observations transient or permanent changes are known in the literature as outliers (sometimes referred to as aberrant or atypical observations) and, depending on their nature, their effects on inference can be substantial. For example, the presence of outliers can increase the estimated variance of the stochastic process, which implies a decrease in the estimated autocorrelations and consequent loss of information about the autocorrelation structure of the process see, for example, Chan [7,8]. Although long-memory models in the presence of outliers has recently been a subject of much interest to some researchers, especially in the areas of economics and finance (Beran [4]; Franses et al. [14]; Tolvi [52]), very few papers have been devoted to this topic in the environmental field. One of the contributions of this paper is to fill this gap. To this purpose, it concatenates long-memory modeling (with more than one fractional parameter), robust estimation and pollution data.

Based on the above discussion, the main goal of this paper is to propose a robust semiparametric estimator for both non-seasonal and seasonal long-memory parameters in SARFIMA models based on a robust autocovariance function estimator. Some model and estimation properties are discussed and a small sample size investigation is conducted to show the method's performance under contaminated and uncontaminated SARFIMA models. In addition, the method is applied to model and forecast SO_2 concentrations, measured at the air quality automatic monitoring network (AQAMN) of the Greater Vitória Region (GVR), ES, Brazil. This series exhibits seasonality, long memory behavior and high levels (or large peaks) of SO_2 concentrations. These observations may produce sample densities with heavy tails and it can strongly influence sample functions such as the standard mean, the covariance and the periodogram. Since the estimation of time series models is connected with these sample functions, the final estimated model can be strongly affected by the large peaks of concentration.

This paper is organized as follows. Section 2 introduces the model, discusses its properties and summarizes its estimation methods. Section 3 presents a simulation study and Section 4 deals with the analysis, modeling and forecasting of SO_2 concentrations. Some conclusions are drawn in Section 5.

2. Model definitions and a robust parameter estimation method

2.1. Model definition and properties

Let $X_t \equiv \{X_t\}_{t\in\mathbb{Z}}$ be a zero-mean process with infinite MA representation $X_t = \Psi(B)\epsilon_t$ where

$$\Psi(B) = \psi(B)\nabla^{\alpha_1,\ldots,\alpha_L}_{\omega_1,\ldots,\omega_L}(B) = \psi(B)\prod_{\iota=1}^{L}(1 - 2\cos\omega_\iota.B + B^2)^{\alpha_\iota}, \tag{1}$$

$\{\epsilon_t\}_{t\in\mathbb{Z}}$ is a white noise with $\mathbb{E}(\epsilon_t) = 0$; $\mathrm{Var}(\epsilon_t) = \sigma^2_\epsilon$, $\omega_\iota \in (-\pi, \pi]$; B is the backward shift operator satisfying $B^k Y_t = Y_{t-k}$ for any process $\{Y_t\}_{t\in\mathbb{Z}}$, $\alpha_\iota = -d_\iota$, $d_\iota \in \mathbb{R}$ $(d_\iota > -1)$, $(d_\iota \neq 0)$ and d_ι is defined as the fractionally differencing parameter. The following assumptions are made for the power expansion of $\Psi(B)$:

(A1) The function $\psi(z)$ is analytic inside and on the unit circle, $|z| \leq 1$ for all $z \epsilon \mathbb{C}$, that is, the coefficients ψ_j, $j = 0, 1, \ldots$, of $\psi(z)$ satisfy $\sum_{j=0}^{\infty}\psi_j^2 < \infty$.

(A2) $|d_\iota| < \frac{1}{2}\ \forall\ \iota = 1, \ldots, L$.

Under assumptions A1 and A2, X_t is a stationary process with spectral density of the form:

$$f_X(\omega) = g_\psi(\omega)|\omega|^{-2d}\prod_{\iota=1}^{L}\prod_{j=1}^{\xi_\iota}|\omega - \omega_{\iota j}|^{-2d_\iota}, \tag{2}$$

where $\omega \in (-\pi, \pi]$ and $g_\psi(\omega)$ is a continuous function, bounded above and away from zero and $\omega_{\iota j} \neq 0$ are poles for $j = 1, \ldots, \xi_\iota$, $\iota = 1, \ldots, L$. The autocovariance function of X_t behaves like $\gamma_X(h) \sim Kj^{2d_\iota - 1}\cos(j\omega)$ as $h \to \infty$ and K is a constant that does not depend on h. See, for example, Giraitis and Leipus [16], Palma [33], Arteche [2], Arteche and Robinson [3], Reisen et al. [45] and references therein.

For suitable choices of the fractionally differencing parameters d_ι, $\iota = 1, \ldots, L$, X_t may have a finite number of zeros or singularities of order $d_1, \ldots, d_L$ on the unit circle which allows the modeling of long and short memory data containing seasonal periodicities.

A stationary long-memory time series with memory parameter $\varphi \in (-0.5, 0.5)$ has autocovariance function $\gamma(h)$ and spectral density $f(\omega)$ satisfying, respectively,

(A3)

$$\gamma(h) \sim h^{2\varphi-1}L_1(h), \quad as \quad h \longrightarrow \infty, \tag{3}$$

(A4)

$$f(\omega) \sim |\omega|^{-2\varphi}L_2(\omega), \quad as \quad \omega \longrightarrow 0, \tag{4}$$

where L_1 and L_2 are slowly varying functions, L_1 at infinity and L_2 at zero (see, for example, Taqqu [51]). When $0 < \varphi < 1/2$, the autocovariance is not absolutely summable and the spectral density becomes unbounded at zero frequency. The properties given by Assumptions 3 and 4 can also be extended for any frequency $\omega \in (-\pi, \pi]$, as here discussed (see Remark 2 for the spectral density function) and in Reisen et al. [45] among others.

Now, let $\psi(B) = \frac{\theta_p(B)}{\phi_q(B)}$ where $\theta_q(z)$ and $\phi_p(z)$ are polynomials with order p and q, respectively. It is assumed that these polynomials have no common roots and satisfy the conditions $\phi_p(z), \theta_q(z) \neq 0$, for all $z \epsilon \mathbb{C}$, such that $|z| \leq 1$. Then, X_t becomes the ARUMA$(p, d_1, ..., d_L, q)$ model introduced by Giraitis and Leipus [16]. Let now d_ι, $\iota = 1, \ldots, L$, satisfying the following assumption,

(A5)

$$|d_\iota| < \begin{cases} 1/2, & 0 < \omega < \pi, \\ 1/4, & \omega = 0, \pi. \end{cases}$$

Under the conditions on $\phi_p(z)$ and $\theta_q(z)$ and Assumption 5, Theorem 2 in Giraitis and Leipus [16] states that the process X_t is causal, invertible and has the unique stationary solution

$$X_t = \frac{\theta_q(B)}{\phi_p(B)}\nabla^{-d_1,\ldots,-d_L}_{\omega_1,\ldots,\omega_L}\epsilon_t = \sum_{j=0}^{\infty}\psi_j\nabla^{-d_1,\ldots,-d_L}_{\omega_1,\ldots,\omega_L}\epsilon_{t-j}. \tag{5}$$

Additionally, the authors show that the spectral density is given by

$$f_X(\omega) = |\theta_q(e^{-i\omega})|^2\phi_p(e^{-i\omega})|^{-2}f_\nabla(\omega) \tag{6}$$

where $f_\nabla(\omega) = \frac{\sigma^2_\epsilon}{2\pi}|\nabla^{-d_1,\ldots,-d_L}(e^{-i\omega})|^2$ and the autocovariance is

$$\gamma_X(h) = \sum_{j=1}^{L}|\theta_q(e^{-i\omega_j})|^2\phi_p(e^{-i\omega_j})|^{-2}a_j \mid h|^{2d_J-1}(\cos h\omega_j + o(1)) \quad \text{as} \quad h \to \infty, \tag{7}$$

where a_j, $j = 1, \ldots, L$, are constants that depend on d_j, that is, they do not depend on h.

Remark 1. If there exists at least one $d_\iota > 1/2$, the process in Eq. (5) is non-stationary and therefore the spectral representation given in Eq. (6) does not exist. Nevertheless, the model is still adequate to adjust time series with long-memory and seasonality. There is a large amount of papers in the literature related to the estimation of non-stationary ARFIMA models (see, for example, Hurvich and Ray [21] and Olbermann et al. [32]).

The SARFIMA model is a particular case of the ARUMA process, that is, it is intrinsically related to the model and assumptions described before. As previously discussed, the SARFIMA model has been widely studied, theoretically and empirically, and applied to a variety of real data set (see, for example, Ray [38], Marques [30], Hassler [18], Reisen et al. [42,43], Arteche and Robinson [3], Palma and Chan [34], among others). Reisen et al. [45] studied the SARFIMA process which encompasses two seasonal fractional and short memory parameters by deriving the model and the asymptotic properties of the semiparametric ordinary least square estimator (OLS). Under some conditions, they showed that the fractional OLS estimators are asymptotically Normally distributed. In their study, the theoretical properties were investigated for finite sample sizes under different scenarios. In this direction, hereafter, the SARFIMA model considered is a particular case of the one discussed by Reisen et al. [45], that is, the SARFIMA model with two fractional parameters (at zero and seasonal frequencies), but in the context of robustness of the OLS long-memory estimators.

Let now

$$\nabla^{\mathbf{d}} = (1-B)^d(1-B^s)^D, \tag{8}$$

where $\mathbf{d} = (d, D)'$ is the memory parameter vector; d and D are the fractional parameters at the zero (or long-run) and seasonal frequencies, respectively, satisfying Assumption 5 and $s \in \mathbb{N}^* = \mathbb{N} - \{0\}$ is the seasonal length. $D = 0$ implies that the process does not have seasonal poles. The process $X_t \equiv \{X_t\}_{t\in\mathbb{Z}}$ is now defined as a zero-mean SARFIMA process satisfying

$$X_t = \psi(B)\nabla^{-\mathbf{d}}\epsilon_t = \frac{\theta_q(B)}{\phi_p(B)}\nabla^{-\mathbf{d}}\epsilon_t, \tag{9}$$

where the process ϵ_t, the fractional vector $\mathbf{d} = (d, D)'$ and the coefficients $\phi_p(B)$ and $\theta_q(B)$ satisfy the conditions previously stated.

In addition, the fractional filters are

$$(1 - B^k)^x = \sum_{j=0}^{\infty} \binom{x}{j} \left(-B^k\right)^j, k = 1, s, \text{ and } x = d, D,$$

where

$$\binom{x}{j} = \frac{\Gamma(x+1)}{\Gamma(j+1)\Gamma(x-j+1)},$$

and $\Gamma(\cdot)$ is the well-known Gamma function.

Remark 2. If $|d+D| < 1/2$, $|d| < 1/2$ and $|D| < 1/2$, X_t is a stationary and invertible process. These conditions are derived by Assumption 5. At seasonal frequencies $\omega_s \in (0, \pi]$, the spectral density becomes unbounded and behaves as

$$f(\omega + \omega_s) \sim C_1|\omega|^{-2D}, \quad \omega \to 0, \tag{10}$$

and at the zero frequency,

$$f(\omega) \sim C_2|\omega|^{-2(d+D)}, \quad \omega \to 0, \tag{11}$$

where $0 < C_1, C_2 < \infty$ (see Proposition 1 in Reisen et al. [45]). From the above, it can be seen that the process has spectral density with poles at zero and seasonal frequencies.

If, in addition to long-memory and periodicity features, X_t also presents outliers, it is necessary to build a model estimation method that encompasses this type of characteristic. To achieve this, the fractional parameter estimation procedure here suggested concatenates the methods given in Reisen et al. [45], Lévy-Leduc et al. [24–26] and Molinares et al. [31] to estimate the fractional parameters of the model presented in Eq. (9). The robust estimation method is discussed in the next sub-section.

2.2. A robust semiparametric estimator for the vector $\mathbf{d}$ in the SARFIMA model

Let $\{X_1, \ldots, X_n\}$ be a sample from the process in Eq. (9). The periodogram function of X_t is given by

$$I_{n,X}(\omega_j) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} X_t e^{i\omega_j t}\right|^2, \tag{12}$$

where $\omega_j = \frac{2\pi j}{n}$, $j = 1, \ldots, (\frac{n}{2} - 1)$, are the Fourier frequencies. Since X_t is a stationary process, $I_{n,X}(\omega_j)$ can be also written as follows:

$$I_{n,X}(\omega_j) = \frac{1}{2\pi}\sum_{k=-(n-1)}^{n-1} \hat{\gamma}_X(k)e^{-i\omega_j k}, \tag{13}$$

where $\hat{\gamma}_X(k)$ is the sample autocovariance function of $\{X_1, \ldots, X_n\}$.

An alternative robust spectral estimator for the ARFIMA process was proposed in Molinares et al. [31] which replaces the classical sample autocovariance $\hat{\gamma}_X(k)$ in Eq. (13) by the robust autocovariance function given in Ma and Genton [29]. The main asymptotic results of the robust autocorrelation function (ACF) for long-memory processes are discussed in Lévy-Leduc et al. [26]. Here, as previously mentioned, the robust ACF estimator (Ma and Genton [29])

will be used to obtain the estimates of the parameters d and D in Model (9). Therefore, the estimation approach here proposed is an extension to the methods discussed by Reisen et al. [45].

For a sample $\nu = (\nu_1, \nu_2, \ldots, \nu_{n'})'$, Rousseeuw and Croux [46,47] suggested the scale robust estimator $Q_{n'}(.)$, which is based on the τth order statistic of $\binom{n'}{2}$ distances $\{|\nu_j - \nu_k|,\ j < k\}$, and can be written as

$$Q_{n'}(\nu) = c \times \{|\nu_j - \nu_k|;\ j < k\}_{(\tau)}, \tag{14}$$

where c is a constant used to guarantee consistency ($c = 2.2191$ for the normal distribution) and $\tau = \left\lfloor \frac{\binom{n'}{2}+2}{4} \right\rfloor + 1$.

Based on the scale robust estimator $Q_{n'}(\nu)$, Ma and Genton [29] proposed the following robust sample autocovariance function

$$\widehat{\gamma}_{Q_{n',y}}(h) = \frac{1}{4}\left[Q^2_{n'-h,y}(\mathbf{u}+\mathbf{v}) - Q^2_{n'-h,y}(\mathbf{u}-\mathbf{v})\right], \tag{15}$$

where $\mathbf{u}$ and $\mathbf{v}$ are vectors containing the initial $n'-h$ and the final $n'-h$ observations, respectively, of a time series sample $Y_1, \ldots, Y_{n'}$ of a process $Y_{t\in\mathbb{Z}}$ with absolutely summable autocovariance function, that is, $\sum_{h=0}^{\infty}|\gamma_y(h)| < \infty$ where $\gamma_y(h)$ is the autocovariance of Y_t at the lag h .

As pointed out by the authors, the robust estimator of the autocorrelation function could be obtained by dividing $\widehat{\gamma}_{Q_{n',y}}(h)$ in Eq. (15) by the product of $Q_{n',y}(\mathbf{u})$ and $Q_{n',y}(\mathbf{v})$. However, this would not be a natural autocorrelation estimator because it would not be bounded between -1 and 1. Thus, a highly robust autocorrelation estimator can be computed by

$$\widehat{\rho}_{Q_{n',y}}(h) = \frac{Q^2_{n'-h,y}(\mathbf{u}+\mathbf{v}) - Q^2_{n'-h,y}(\mathbf{u}-\mathbf{v})}{Q^2_{n'-h,y}(\mathbf{u}+\mathbf{v}) + Q^2_{n'-h,y}(\mathbf{u}-\mathbf{v})}. \tag{16}$$

It can be shown that $|\widehat{\rho}_{Q_{n',y}}(h)| \leq 1$ for all h. As previously mentioned, the asymptotic properties of $\widehat{\gamma}_{Q_{n',y}(h)}$ when the time series Y_t has short and long-memory properties are discussed in Lévy-Leduc et al. [24–26]. Here, some of their results are addressed in the following remarks.

Remark 3. Under the assumption that $Y_{t\in\mathbb{Z}}$ follows a Gaussian long-memory process with memory parameter $0 < \varphi < 1/2$ (see Assumptions 3 and 4), Lévy-Leduc et al. [26] showed asymptotic results for $Q_{n',y}(.)$ and $\widehat{\gamma}_{Q_{n',y}}(h)$. In particular, the authors demonstrated that for $1/4 < \varphi < 1/2$, the robust autocovariance estimator $\widehat{\gamma}_{Q_{n',y}}(h)$ has the same asymptotic behavior as the classical autocovariance estimator $\hat{\gamma}_y(h)$. In this case, there is no loss of efficiency. For $0 < \varphi < 1/4$, $\widehat{\gamma}_{Q_{n',y}}(h)$ has the same rate of convergence of $\hat{\gamma}_y(h)$, but with different variances. The standard Gaussian ARFIMA(p, φ, q) is one particular case of the model discussed by the authors.

Remark 4. Note that, although the Gaussian distribution is required to obtain the asymptotic distribution properties of $Q_{n',y}(.)$ and $\widehat{\gamma}_{Q_{n',y}}(h)$, the finite sample size investigation given in Lévy-Leduc et al. [26] showed that the estimation method also performs well under non Gaussian observations, that is, the robust autocovariance estimator does not seem to be affected by the skewness of the data. Additionally, Molinares et al. [31], Lévy-Leduc et al. [25,26] and Sarnaglia et al. [49], using finite sample size, discuss the finite sample robustness property of $\widehat{\gamma}_{Q_{n',y}}(h)$ under different scenarios of the data contaminated with additive outliers, that is, observations which can produce skewness in the series. Their investigations strongly suggest the use of $\widehat{\gamma}_{Q_{n,y}}(h)$ in this context. A thorough search of the relevant literature on this topic indicated that, under skewness distribution, the asymptotic property of $\widehat{\gamma}_{Q_{n,y}}(h)$ is very difficult to be obtained and it is still an open problem.

Remark 5. From Eq. (7), it can be seen that the autocovariances of the process X_t, given in Eq. (9), show an asymptotic slow decay typical of Eq. (3) but with oscillations that depend on the frequency ω, as shown in Proposition 1 of Reisen et al. [45]. Additionally, from Remark 2, the process has spectral density with poles at zero and seasonal frequencies. Under the assumption that the innovations ϵ_t are from a Gaussian distribution, the SARFIMA model in Eq. (9) satisfies the model conditions given in Lévy-Leduc et al. [26] (see Assumptions 3 and 4) and, therefore, the asymptotic results for the robust autocovariance discussed in Lévy-Leduc et al. [26] for the ARFIMA model are also valid for samples from the SARFIMA process (see, Remark 3). The finite sample properties discussed in Remark 4 are also expected for the model here studied (see the finite sample size investigation analysis in Section 3).

Based on the previous discussion and on Molinares et al. [31], for the sample $X_1, \ldots, X_n$ of Model (9), a robust spectral estimator can be computed as follows

$$I_{Q_n,X}(\omega) = \frac{1}{2\pi} \sum_{|h|<n} \kappa(h) \widehat{\gamma}_{Q_n,X}(h) \cos(h\omega), \tag{17}$$

where $\kappa(h)$ is defined as

$$\kappa(h) = \begin{cases} 1, & |h| \leq \xi, \\ 0, & |h| > \xi. \end{cases} \tag{18}$$

$\kappa(h)$ is a particular case of the *lag window* functions used in classical spectral theory to obtain a consistent spectral estimator, and ξ is the truncation point which is a function of n, say $\xi = G(n)$, where $G(n)$ must satisfy $G(n) \to \infty, n \to \infty$, with $\frac{G(n)}{n} \to 0$. $G(n)$ is usually chosen to be $G(n) = n^{\beta}$, where $0 < \beta < 1$ (see, e.g., Priestley [36, pp. 433–437]). In addition, the robust ACF estimator given in Eq. (16) does not have the same finite-sample properties as the classical one. For large h, the number of observations in the calculation of $\widehat{\gamma}_{Q_n,X}(h)$ is very small and, consequently, $\widehat{\rho}_{Q_n,X}(h)$ becomes very unstable. Therefore, the bandwidth ξ will avoid these undesirable covariance estimates in the calculation of the estimator given in Eq. (17). See Assumption 3 in Molinares et al. [31] for more discussion on the choice of $G(n)$. Note that, similar to the classical spectral estimation theories, other *lag window* functions can be used to obtain a robust spectral estimator, as discussed in Molinares et al. [31].

To estimate $\mathbf{d} = (d, D)'$ for the SARFIMA process in Eq. (9), Reisen et al. [45] suggest the ordinary least squares estimator (OLS) $\hat{\mathbf{d}}_{CL} = (\hat{d}_{CL}, \hat{D}_{CL})'$ computed from the approximated multiple linear regression equation

$$\log I_{n,X}(\omega_{kj}) \cong a_0 - D \log \left[2 \sin\left(\frac{s\omega_{kj}}{2}\right)\right]^2 - d \log \left[2 \sin\left(\frac{\omega_{kj}}{2}\right)\right]^2 + u_{kj}, \tag{19}$$

where a_0 is a constant and

$$u_{kj} = \log \frac{I_{n,X}(\omega_{kj})}{f_X(\omega_{kj})} - \mathbb{E}\left[\log \frac{I_{n,X}(\omega_{kj})}{f_X(\omega_{kj})}\right].$$

The frequencies $\omega_{kj}, k = 0, 1, \ldots, [\frac{s}{2}], 1 \leq j \leq M$, are defined as

$$\omega_{kj} = \begin{cases} \dfrac{2\pi k}{s} + \dfrac{2\pi j}{n}, & k = 0, \\ \dfrac{2\pi k}{s} \pm \dfrac{2\pi j}{n}, & k = 1, \ldots, [\dfrac{s}{2}] - 1. \end{cases} \tag{20}$$

$$\omega_{[\frac{s}{2}]j} = \begin{cases} \dfrac{2\pi[\frac{s}{2}]}{s} - \dfrac{2\pi j}{n}, & s \text{ even}; \\ \dfrac{2\pi[\frac{s}{2}]}{s} \pm \dfrac{2\pi j}{n}, & s \text{ odd}. \end{cases}$$

In the above equations, $M = M(n)$ is the bandwidth that has to satisfy

(A6)

$$\left(\frac{M}{n}\right)^{\alpha} \log M + \frac{1}{M} \to 0\,, \quad \text{as } n \to \infty\,,$$

for some $\alpha > 0$. One suggestion is to use an α that avoids overlapping frequencies, for example, $M < \frac{n-1}{2s}$ (see, also, Remark 7).

Under some assumptions which also includes Gaussian innovations, Reisen et al. [45] establish that

$$\sqrt{M}(\hat{\mathbf{d}}_{CL} - \mathbf{d}) \to \mathcal{N}\left(W^{-1}b, \frac{\pi^2}{6} W^{-1}\right)\,, \tag{21}$$

where $\mathbf{b}$ and $\mathbf{W}$ are a vector and a 2×2 matrix of constants, respectively. See, Theorems 1 and 2 in Reisen et al. [45].

In this work, a robust OLS estimator $\hat{\mathbf{d}}_R = (\hat{d}_R, \hat{D}_R)'$ is proposed replacing $I_{n,X}(\omega_{kj})$ in Eq. (19) by $I_{Q_n,X}(\omega_{kj})$, given in Eq. (17). The choice of the bandwidths $M = M(n)$ will be based on Assumption 6 under restrictions of Eq. (20).

Remark 6. Note that, although the asymptotic distribution of $\hat{\mathbf{d}}_R$ is still an open problem, the consistency and distribution properties of the estimators, finite sample properties of the estimators of the long-memory parameter and the sample autocovariance functions in the ARFIMA model are investigated by Molinares et al. [31] and Lévy-Leduc et al. [26], respectively. These works together with the simulation results displayed in the following section for the SARFIMA process support the use of the robust function Eq. (17) as an alternative spectral estimator to obtain robust fractional parameter estimates in a real time series that presents long-memory and additive outliers features. The $\hat{\mathbf{d}}_R$ estimator is implemented in R-project [37] and the code can be obtained upon request.

Remark 7. Note that, if AR and/or MA coefficients are introduced in the model these short-memory parameters lead to biased estimates and the bias (positive or negative) will depended on the size of the bandwidth M and on the values of the AR/MA parameters. This issue is well-documented in Reisen et al. [45], Reisen [39] and Hurvich and Ray [21]. See, also, Hassler [18] for the flexible SARFIMA model.

3. A simulation study

In this section, the finite sample performance of the OLS robust fractional parameter estimator ($\hat{\mathbf{d}}_R = (\hat{d}_R, \hat{D}_R)'$) is investigated through Monte Carlo experiments for SARFIMA models, with $p = q = 0$ and $s = 12$, and with i.i.d innovations from a $N(0, 1)$ distribution. For comparison purposes, the classical OLS fractional estimator ($\hat{\mathbf{d}}_{CL} = (\hat{d}_{CL}, \hat{D}_{CL})'$), which is based on the classical periodogram (Reisen et al. [45]) is also considered in the simulation study. Since the simulated models here do not have AR and/or MA components, the OLS estimates were computed based on the bandwidth $M = [(n-2s)/2s]$ in order to avoid overlapping frequencies (see, also, Assumption 6 and Remark 7). For the robust spectral estimator (Eqs. (17) and (18)), the bandwidth was $G(n) = n^{0.7}$ (see, Molinares et al. [31]). The sample size is $n = 1000$ and the mean and standard deviation were calculated over 1000 replications. Other scenarios with smaller sample sizes (for example, $n = 500, s = 3$) were also investigated resulting in similar conclusions (they are available upon request).

Fig. 1 shows the box-plots with the results of both estimators for series generated without outliers (no contamination). It can be seen that, in general, both methods perform similarly, that is, under the scenario of a non-contaminated time series both estimation methods lead to comparable results, with estimates close to the real values of d and D. However, the robust estimates have slightly smaller variation compared with the classical ones (see, also, Molinares et al. [31] and Lévy-Leduc et al. [26] for ARFIMA models).

Fig. 2 and 3 present the estimated densities of the classical and robust estimators for SARFIMA models with $d = D = 0.1$ and $d = 0.3$ and $D = 0.1$, respectively. Although, as previously stated, the asymptotic distribution of $\hat{\mathbf{d}}_R$ was not yet proved, it can be seen that the standardized estimates are fairly close to the density of the $N(0, 1)$ distribution. Therefore, this simple simulation study can give some support for the theoretical discussion presented in the previous remarks and the use of the proposed estimation method in the context of this paper, that is, in time series with long-memory, seasonality and outliers.

The discussion of the performance of the robust autocovariance estimate under non-Gaussian distribution made in Remark 4 is also valid for the parameter estimation. Fig. 4 displays the densities of the estimates when the innovations follow a Student's t-distribution with 3 degrees of freedom. The performance of the estimates was similar to the Gaussian case. See, also, Lévy-Leduc et al. [26] (Section 3.3) for non-Gaussian observations.

To investigate the robustness property of $\hat{\mathbf{d}}_R = (\hat{d}_R, \hat{D}_R)'$ in a time series under outliers, the SARFIMA model contaminated with additive outliers was simulated according to the model structure given in Lévy-Leduc et al. [26] and Molinares et al. [31] and this is summarized below.

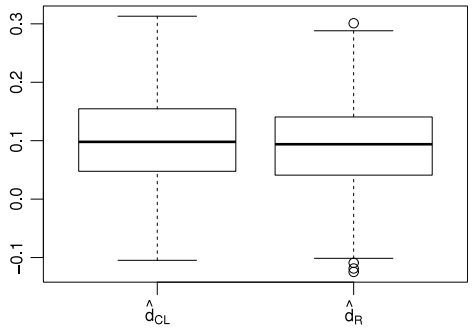
Let now Y_t be defined by

$$Y_t = X_t + \varpi W_t, \tag{22}$$

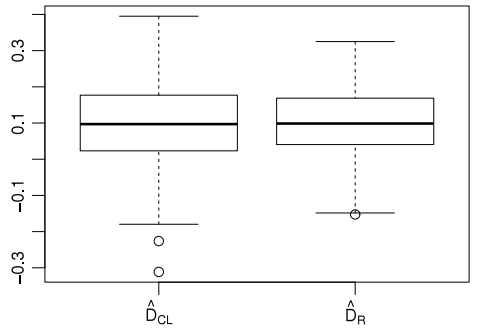
where the parameter ϖ represents the magnitude of an additive outlier and W_t is a random variable with probability distribution

$$P(W_t = -1) = P(W_t = 1) = \delta/2 \text{ and } P(W_t = 0) = 1 - \delta,$$

where $\mathbb{E}[W_t] = 0$ and $\mathbb{E}[W_t^2] = \text{Var}(W_t) = \delta$. Note that Eq. (22) is based on the parametric models proposed by Fox [13]. W_t is the product of *Bernoulli* (δ) and *Rademacher* random variables; the latter equals 1 or -1, both with probability $1/2$. X_t and W_t are independent random variables.

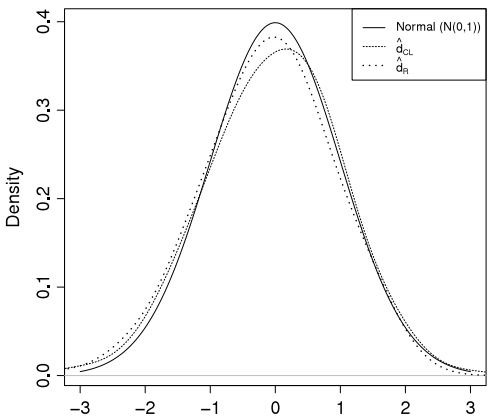


(a) Estimates of d (classical $\hat{d}_{CL}$ and robust $\hat{d}_R$).

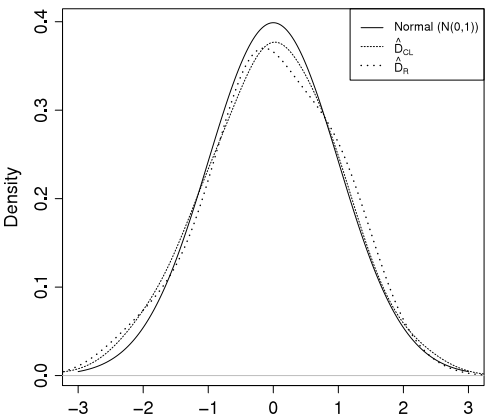


(b) Estimates of D (classical $\hat{D}_{CL}$ and robust $\hat{D}_R$).

Fig. 1. Box-plots of the estimates $\hat{d}_{CL}$, $\hat{d}_R$, $\hat{D}_{CL}$ and $\hat{D}_R$ for the SARFIMA model with $p = q = 0$, $d = D = 0.1$ and $s = 12$. Non-contaminated series.

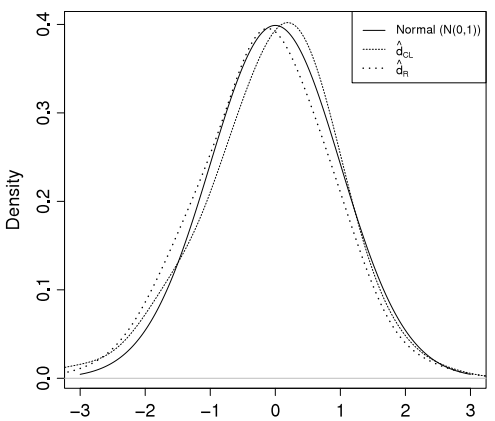


(a) Empirical densities of $N(0, 1)$, $\hat{d}_{CL}$ and $\hat{d}_R$.

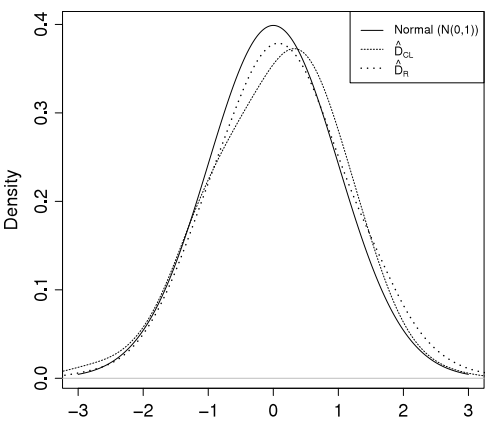


(b) Empirical densities of $N(0, 1)$, $\hat{D}_{CL}$ and $\hat{D}_R$.

Fig. 2. Empirical densities of the $N(0, 1)$ and of the standardized estimates for the SARFIMA with $p = q = 0$, $d = D = 0.1$ and $s = 12$.

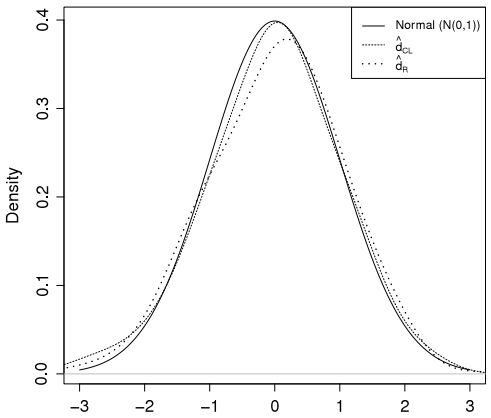


(a) Empirical densities of $N(0, 1)$, $\hat{d}_{CL}$ and $\hat{d}_R$.

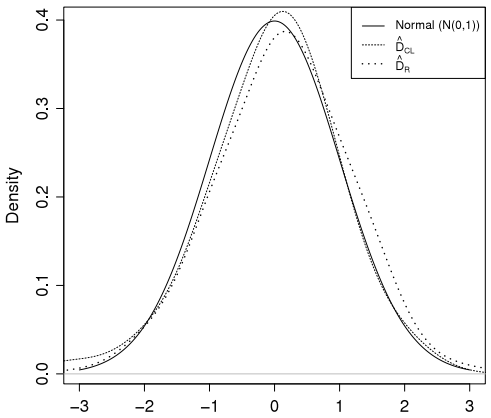


(b) Empirical densities of $N(0, 1)$, $\hat{D}_{CL}$ and $\hat{D}_R$.

Fig. 3. Empirical densities of the $N(0, 1)$ and of the standardized estimates for the SARFIMA model with $p = q = 0$, $d = 0.3$, $D = 0.1$ and $s = 12$.

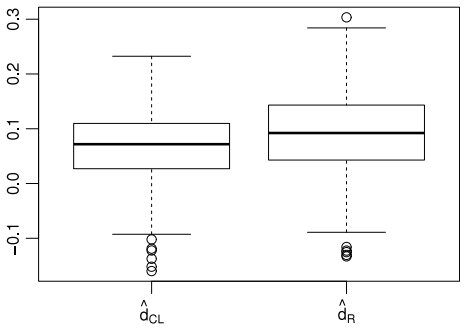


(a) Empirical densities of $N(0, 1)$, $\hat{d}_{CL}$ and $\hat{d}_R$.

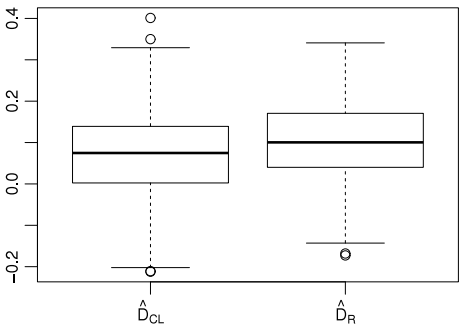


(b) Empirical densities of $N(0, 1)$, $\hat{D}_{CL}$ and $\hat{D}_R$.

Fig. 4. Empirical densities of the standardized estimates for the SARFIMA model with innovations generated from a *t-student* distribution (3 d.f.) and $p = q = 0$, $d = 0.1$, $D = 0.1$ and $s = 12$.



(a) Estimates of d.



(b) Estimates of D.

Fig. 5. Box-plots of the estimates $\hat{d}_{CL}$, $\hat{D}_{CL}$, $\hat{d}_R$ and $\hat{D}_R$ for the SARFIMA model with $p = q = 0$, $d = D = 0.1$ and $s = 12$, for series with outliers.

Fig. 5 presents the box-plots for the classical and robust estimators considering the series with outliers ($\varpi = 15$ and $\delta = 0.05$). As can be seen from the box-plots, the classical estimator is clearly affected by additive outliers while the robust one keeps almost the same picture of the non-contaminated scenario. This simple investigation leads to analogous conclusions given in Lévy-Leduc et al. [26] and Molinares et al. [31], when the generated process follows an ARFIMA model, that is, the classical OLS fractional estimator is completely influenced by the outliers while, in general, the robust one is not. In the non-Gaussian series, the methods displayed a similar performance and the results are available upon request.

4. An application to SO_2 pollutant

The daily average SO_2 concentration is expressed in $\mu g/m^3$ and was measured at the Air Quality Automatic Monitoring Network (AQAMN) of Cariacica, which belongs to the Metropolitan area of the Great Vitória Region (GVR) – ES – Brazil. GVR is comprised of seven cities with a population of about 1.7 million inhabitants in an area of 2331 km^2. The region is situated along the South Atlantic coast of Brazil (latitude 20°19S, longitude 40°20W) and has a warm tropical climate, with average temperatures ranging from 24 °C (Celsius) to 30 °C.

The raw series has a sample size of 1826 daily observations (Fig. 6), measured from January 1st 2005 to December 31st 2009. The maximum concentration is generally observed in the winter months (southern hemisphere) from July to September. It can be seen that the series has some high concentration of SO_2 in different points in time. As previously

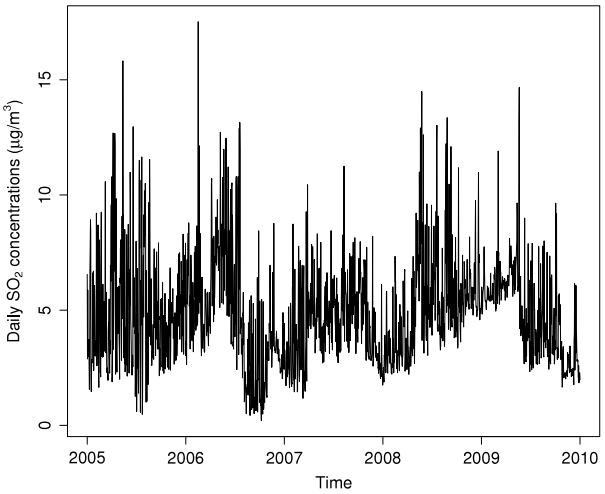


Fig. 6. Daily SO_2 concentrations ($\mu g/m^3$) from 2005/01/01 to 2009/12/31.

mentioned, these large peaks can be viewed here as outliers, since their values can provoke serious damage to the statistical functions, such as the mean and the standard deviation and, therefore, may affect the correlation structure of the series, leading to misleading results.

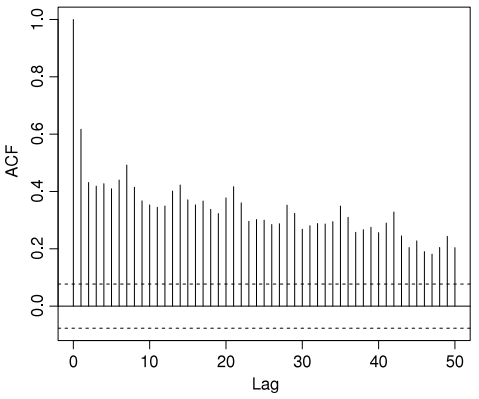
Since the variability in the series did not seem to be stable, a log transformation was used ($Z_t = log(X_t)$). Moreover, the series was divided into two parts: learning and prediction sets. The 1626 observations from January 1st 2005 to June 14th 2009 were considered as learning set and the remaining 200 observations were kept for the forecasting study.

The potential effect of the large concentrations on the sample summary functions is now addressed The sample autocorrelation (ACF), the partial autocorrelation (PACF) and the periodogram functions of Z_t are shown in Fig. 7. These plots indicate possible seasonal behavior with a period equal to seven, which is an expected result since the data are daily mean levels.

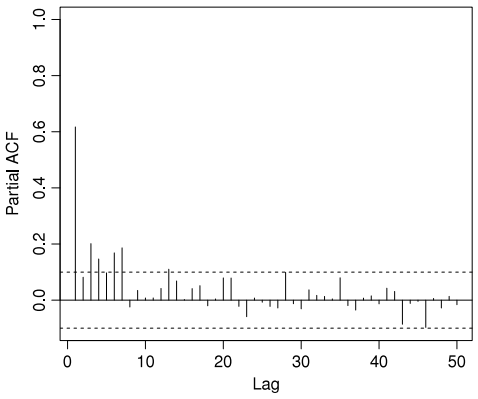
The robust ACF, PACF and periodogram functions are displayed in Fig. 8(a), 8(b) and 8(c), respectively, to compare them with the classical ones in order to examine whether there is any effect of the large peaks on these functions. As can be seen from these figures, the high levels of concentration reduce the size of the classical ACF and PACF functions while increase the peaks of the periodogram; that is, the classical periodogram across the frequencies close to zero are much higher than the robust ones. In particular, the sample ACFs are smaller than the robust ones, for instance, for lags $h = 1, 3, 5, 10$, the robust autocorrelations were equal to 0.72, 0.49, 0.48 and 0.41, while the classical autocorrelations were equal to 0.62, 0.42, 0.41 and 0.35, respectively. Such a behavior was theoretically justified in Corollary 1 of Molinares et al. [31]. Note that the long memory property of the series is well observed by looking at the periodogram plots in Figs. 7(c) and 8(c). Both plots indicate high values for the frequencies close to zero.

Additionally, in Figs. 7(d) and 8(d) the log-periodogram is plotted against the log of the frequencies, for classical and robust cases, respectively. The figures also present the ordinary least square estimator of β_i in the model $log[I(\omega_j)] = \beta_0 + \beta_i log(\omega_j)$, where $i = 1$, if classical, $i = 2$, if robust, and $j = 1, \ldots, M$, with $M = 33$ ($\alpha = 0.465$) which satisfies Assumption 6 and avoids the overlapping frequencies. Comparing the intensity of the long memory dependency (the slopes of the regressions in Figs. 7(d) and 8(d)), it can be seen that the dependency is larger for the robust estimator ($|\hat{\beta}_1| < |\hat{\beta}_2|$). However, as the SO_2 concentrations exhibit seasonality, any conclusion about the estimates $\hat{\beta}_i$, $i = 1, 2$, should be taken with care.

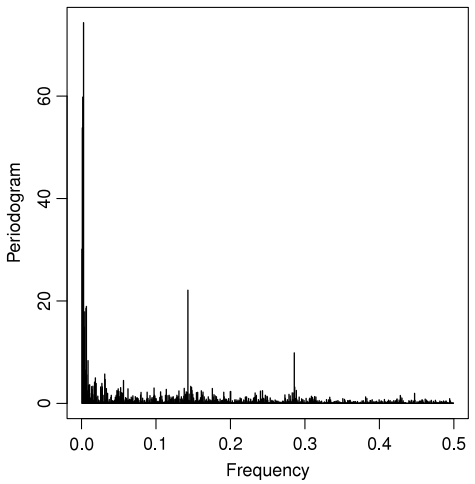
The robust SARFIMA modeling strategy follows similar steps suggested in Hosking [19] and investigated by Reisen [39] and Reisen and Lopes [41]. In semiparametric procedures, the estimation of the model parameters is performed in two steps: firstly, the parameter vector $\mathbf{d}_R$ is estimated based on the procedures presented in Section 2.2. Secondly, the truncated filter $(1 - B)^{\hat{d}_R}(1 - B^s)^{\hat{D}_R}$ is used to filter the observations. This new series is used to estimate the autoregressive and moving average parameters. The fitted models and their accuracy are discussed in the next subsections.



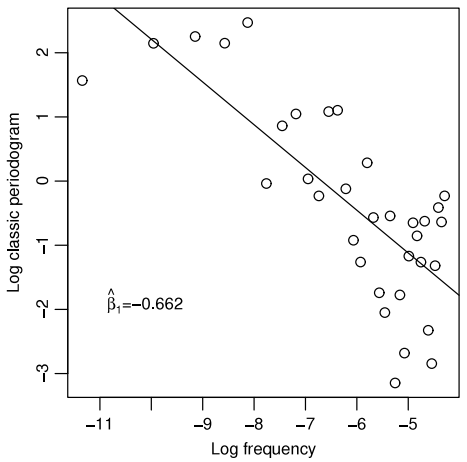
(a) ACF.



(b) PACF.



(c) Periodogram.



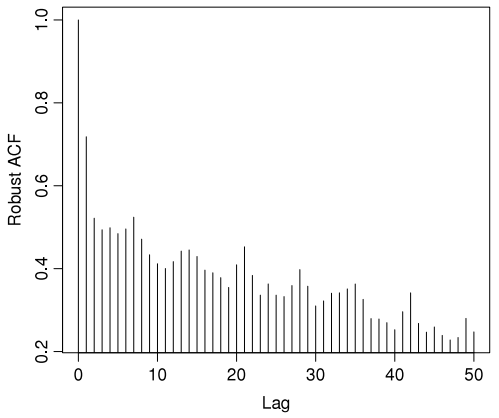
(d) Log(per.) versus log(freq.).

Fig. 7. (a) ACF, (b) PACF and (c) periodogram of Z_t and (d) $log(I_n(\omega_j))$ versus $log(\omega_j)$.

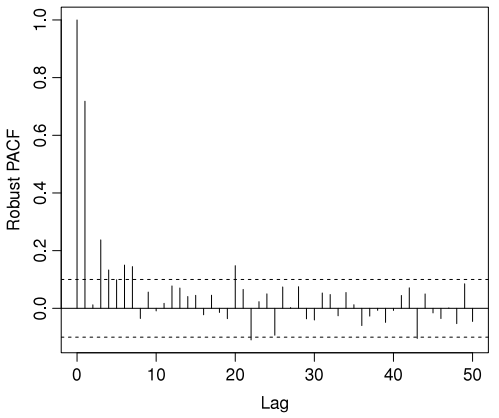
4.1. Adjusted models

The robust and classical estimates of the parameter vector **d** are displayed in Table 1 for different bandwidths M, which corresponds to $0.4 < \alpha < 0.55$, as discussed in Section 2.2. The values in brackets correspond to the standard deviations $(s.d.)$. To compute the robust estimates, the bandwidth in Eq. (18) was fixed as $G(n) = n^{0.7}$. Note that the estimates in Table 1 do not satisfy all the stationary conditions, that is, the values do not satisfy $|d + D| < 0.5$ (see Remarks 1 and 2), but the model still has the mean reverting property in the sense that its cumulative impulse response weights sum to a finite number. In addition, the plot in Fig. 6 suggests a phenomenon of mixture of stationary and nonstationary blocks, which may imply that $0.5 < |\hat{d} + \hat{D}| < 1.0$, although the individual estimates are in the stationary region. However, this possible nonstationarity behavior of the series can be reduced to a stationary time series by differencing the series with a long-memory filter, as is discussed in what follows.

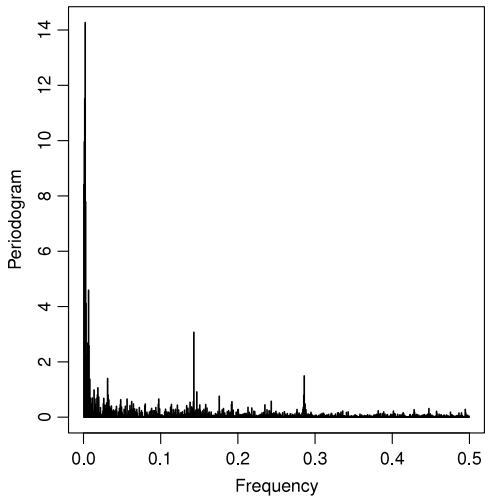
As mentioned in Remark 7, in the semiparametric approach, the choice of estimates depends on the size of the bandwidth. For example, large M leads to more bias in the fractional estimators, when there are short-run components in the model. As expected all estimates in Table 1 using the classical periodogram were lower than those adopting the robust periodogram. This is due to the presence of observations with large peaks in the SO_2 concentrations, as previously discussed. The robust estimates at zero frequency are very stable across the values of M. This is an indication that, if there is a short-memory parameter in the estimated model, it is not large enough to make an impact



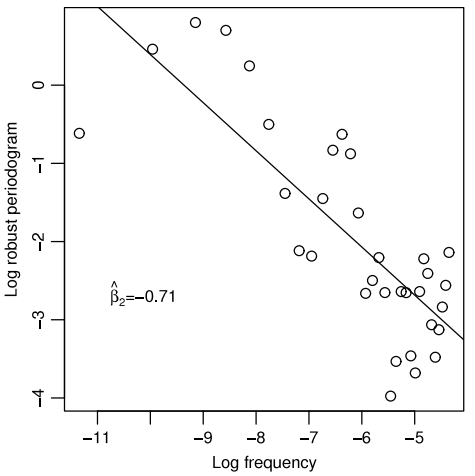
(a) Robust ACF.



(b) Robust PACF.



(c) Robust periodogram.



(d) Log(rob-per.) versus log(freq.).

Fig. 8. (a) Robust ACF, (b) robust PACF and (c) robust periodogram of Z_t and (d) $log(I_{Q_n}(\omega_j))$ versus $log(\omega_j)$ (d).

on the estimates of d (see, also, the ACF plots in Fig. 9). Therefore, $M = 33$ was chosen and thus, using the robust periodogram, $\hat{d}_R = 0.4510$ and $\hat{D}_R = 0.2610$. Note that the robust estimates have smaller $s.d.$ than the standard ones, which is an expected result (see, for example, Molinares et al. [31] and Lévy-Leduc et al. [26] for ARFIMA models).

Now, using the fractional differencing parameter estimates, showed in Table 1 with $M = 33$, the series $\hat{\eta}_t = (1-B)^{\hat{d}_R}(1-B^s)^{\hat{D}_R}Z_t, \quad t = 1, \ldots, 1626$, was obtained. Fig. 9 shows the sample robust ACF and robust PACF functions of $\hat{\eta}_t$.

Based on the sample robust ACF and PACF functions displayed in Fig. 9, some models were considered to fit $\hat{\eta}_t$ and the one which presented the smallest AIC, equal to 1,570.21, was the MA(1) model. Table 2 shows the estimated model. The MA(1) estimate was computed by the Hannan–Rissanen algorithm (see, Brockwell and Davis [5]), in which the classical autocovariance was replaced by the robust one. Although the MA(1) component is adding only a small contribution, the robust SARFIMA$(0, d_R, 1) \times (0, D_R, 0)_7$ was chosen for the SO_2 average data.

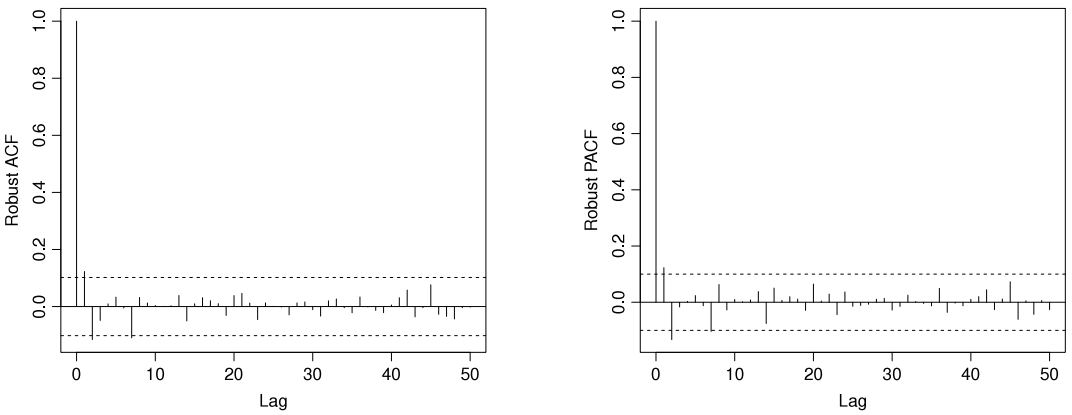
Model adequacy is now addressed (Table 3). The Box–Pierce and Ljung–Box statistics (robust tests) demonstrated that the sample residuals are not time-correlated. In addition, the results indicated that the residuals are not normally distributed, which was an expected result since the original data is skewed to the right. Other classical residual plots were also analyzed and they led to similar conclusions. These plots are available upon request.

Table 1
Estimates of d_R and D_R for different bandwidths (M) for Z_t, using the robust and classical periodogram.

Robust estimates				
M	$\hat{d}_R$	$sd(\hat{d}_R)$	$\hat{D}_R$	$sd(\hat{D}_R)$
25	0.4706	(0.0465)	0.2934	(0.0304)
28	0.4675	(0.0423)	0.2723	(0.0285)
33	0.4510	(0.0398)	0.2610	(0.0280)
38	0.4565	(0.0359)	0.2391	(0.0262)
43	0.4534	(0.0337)	0.2202	(0.0254)
49	0.4511	(0.0307)	0.2100	(0.0240)
57	0.4539	(0.0281)	0.1724	(0.0229)
Classical estimates				
M	$\hat{d}_{CL}$	$sd(\hat{d}_{CL})$	$\hat{D}_{CL}$	$sd(\hat{D}_{CL})$
25	0.4295	(0.0578)	0.2214	(0.0378)
28	0.4303	(0.0552)	0.2080	(0.0372)
33	0.4301	(0.0500)	0.1925	(0.0351)
38	0.4221	(0.0481)	0.1923	(0.0351)
43	0.4099	(0.0451)	0.1952	(0.0341)
49	0.3983	(0.0420)	0.1661	(0.0329)
57	0.3876	(0.0386)	0.1440	(0.0315)

Table 2
Adjusted robust SARFIMA model for SO_2 concentrations.

Parameter	*Estimate*	*sd*
d_R	0.4511	0.0398
D_R	0.2610	0.0228
θ_1	0.0225	0.0145



(a) Robust ACF. (b) Robust PACF.

Fig. 9. Robust ACF and robust PACF of $\hat{\eta}_t$.

An alternative way to demonstrate the usefulness, quality and forecasting performance of the proposed model is to compare it to the standard SARFIMA models. The SARFIMA$(0, d_{CL}, 1) \times (0, D_{CL}, 0)_7$ model (with AIC equal to 1,640.03 and $M = 33$) was chosen to fit the raw data among other candidates and thus, using the classical periodogram, $\hat{d}_{CL} = 0.4301$ and $\hat{D}_{CL} = 0.1925$. The residuals of standard SARFIMA model were right-skewed and uncorrelated. The forecast performance of the robust SARFIMA$(0, d_R, 1) \times (0, D_R, 0)_7$ and the standard SARFIMA$(0, d_{CL}, 1) \times (0, D_{CL}, 0)_7$ models is discussed in the next subsection.

Table 3
Tests for normality* and non correlation (robust tests)**.

Shapiro–Wilk*	Jarque–Bera*	Box–Pierce**	Ljung–Box**
<0.0001	<0.0001	0.8403	0.8404

Note: the p-values correspond to the robust Box–Pierce and robust Ljung–Box test statistics with lag 1. Other lags were tested and they presented similar conclusions.

Table 4
PMSE of the fitted models values to the SO_2 concentration.

Horizon	St. SARFIMA (A)	Rob. SARFIMA (B)	[(A/B)-1] ∗ 100
1	0.0875	0.0857	2.09%
2	0.1088	0.1034	5.27%
3	0.1159	0.1079	7.44%
4	0.1209	0.1126	7.37%
5	0.1241	0.1140	8.86%
6	0.1255	0.1156	8.60%
7	0.1289	0.1161	11.01%
8	0.1392	0.1264	10.15%
9	0.1380	0.1240	11.25%
10	0.1372	0.1236	10.99%

Note: St. = standard; and, Rob. = robust.

4.2. Forecasting investigation

The objective of this section is to verify whether the standard SARFIMA model may lead to less accurate forecasts compared to the robust estimation. As stated before, the observations from June 15th 2009 to December 31st 2009 were discarded from the modeling stage (200 observations) to be used in the out-of-sample forecast study. Forecasts one to ten steps ahead are considered.

To measure the accuracy of the forecasts one to ten steps ahead, the criteria used was the prediction mean square error (PMSE) and the values are displayed in Table 4. From this table, it can be seen that the robust SARFIMA model yields more accurate forecasts than the standard SARFIMA model, especially for long-term forecasts. In addition, the Diebold–Mariano test [12] showed that the forecast of the two models are significantly different and confirmed the superiority forecasting performance of the robust SARFIMA model. For example, for $h = 1$ and 10, the statistical Diebold–Mariano test gave p-values of 0.4×10^{-2} and 0.035, respectively, showing rejection of the null hypothesis in favor of the robust SARFIMA forecasting performance.

Fig. 10 presents a visual analysis of the one-step-ahead forecast values of the robust SARFIMA model, that is, from June 15th 2009 to December 31st 2009. It indicated a reasonably good performance of the model and estimation method proposed here to fit the data. The robust SARFIMA model is able to capture the dynamics of the SO_2 series for one-step-ahead forecasts. These results corroborate the better performance of the fitted robust SARFIMA model to forecast the SO_2 concentration over the standard SARFIMA model, especially for longer lead times.

5. Concluding remarks

This article considered a robust SARFIMA model, with two fractional parameters, in the presence of additive outliers. To estimate the fractional parameters d and D, this paper proposed an estimation procedure which is based on the methods suggested in Reisen et al. [45], Lévy-Leduc et al. [26] and Molinares et al. [31]. Simulation studies demonstrated that the estimation method works well when the series is contaminated with additive outliers. As an example of application, the daily average SO_2 concentration was analyzed to show the usefulness of the proposed methodology. This series presented seasonality, long memory phenomena and some occasional large concentrations. As mentioned in the introduction, the high levels of SO_2 are regarded as potential outliers. Based on the robust estimator of the fractional parameters, the SARFIMA$(0, d_R, 1) \times (0, D_R, 0)_7$ model was used to fit the real data. The results suggested that the residuals were uncorrelated and not normally distributed. The standard estimation method of the SARFIMA model was also considered in the forecasting performance. The PMSE indicated that the robust SARFIMA model had a better accuracy, especially to forecast for long lead times. The robust method is an attractive

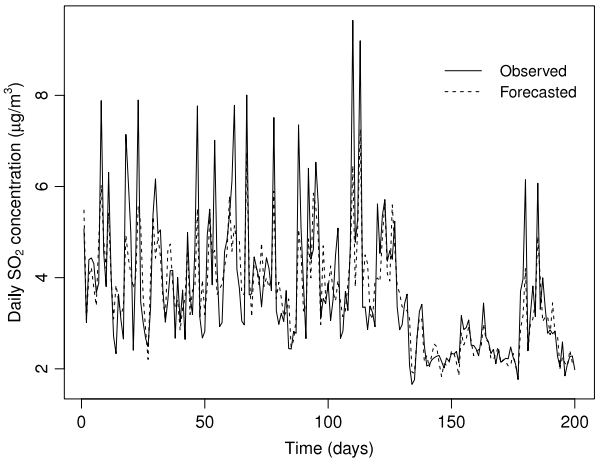


Fig. 10. Forecasted values by the robust SARFIMA model and SO_2 concentrations ($\mu g/m^3$) from June 15th 2009 to December 31st 2009, one-step-ahead.

procedure for estimating the parameters of the SARFIMA model with long memory, seasonality and additive outliers and it can be easily used in application areas. The results in this paper will hopefully stimulate further research on using robust estimation methods and long-memory models to represent and forecast environmental time series.

Acknowledgments

The authors gratefully acknowledge partial financial support from FAPES (67648142/2014), FAPEMIG (PPM00021-14) and CNPq (306234-2015-7, 307099/2014-8). Part of this paper was revised when Prof. Valderio Reisen was visiting the CentraleSupélec (12/2016 to 01/2017). This author is indebted to CentraleSupélec for its financial support. Bovas Abraham was partially supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN-2017-04177). The authors are grateful to the Editor and referees for the time and efforts in providing very constructive and helpful comments that have led to clarify and substantially improve the quality of the paper.

References

[1] S.M. Andersson, B.G. Martinsson, J. Friberg, C.A.M. Brenninkmeijer, A. Rauthe-Schöch, M. Hermann, P.F.J. van Velthoven, A. Zahn, Composition and evolution of volcanic aerosol from eruptions of Kasatochi, Sarychev and Eyjafjallajükull in 2008-2010 based on Caribic observations, Atmos. Chem. Phys. 13 (4) (2013) 1781–1796.
[2] J. Arteche, Semiparametric robust test on seasonal or cyclical long-memory time series, J. Time Ser. Anal. 23 (2002) 251–285.
[3] J. Arteche, P. Robinson, Semiparametric inference in seasonal and cyclical long-memory processes, J. Time Ser. Anal. 21 (2000) 1–25.
[4] J. Beran, On a class of M-estimators for gaussian long-memory models, Biometrika 81 (4) (1994) 755–766.
[5] P.J. Brockwell, R.A. Davis, Time Series: Theory and Methods, second ed., in: Springer Series in Statistics, 2006.
[6] U. Brunelli, V. Piazza, L. Pignato, F. Sorbello, S. Vitabile, Two-days ahead prediction of daily maximum concentrations of SO_2, O_3, PM_{10}, NO_2, CO in the urban area of Palermo, Italy, Atmos. Environ. 41 (14) (2007) 2967–2995.
[7] W. Chan, A note on time series model specification in the presence outliers, J. Appl. Stat. 19 (1992) 117–124.
[8] W. Chan, Outliers and financial time series modelling: a cautionary note, Math. Comput. Simulation 39 (1995) 425–430.
[9] K.S. Chan, H. Tsai, Inference of seasonal long-memory aggregate time series, Bernoulli 18 (4) (2012) 1448–1464.
[10] S. Cheng, K. Lam, Synoptic typing and its application to the assessment of climatic impact on concentrations of sulfur dioxide and nitrogen oxides in Hong Kong, Atmos. Environ. 34 (4) (2000) 585–594.
[11] M.A. Delgado, P.M. Robinson, Optimal spectral bandwidth for long memory, Statist. Sinica 6 (1996) 97–112.
[12] F.X. Diebold, R.S. Mariano, Comparing predictive accuracy, J. Bus. Econom. Statist. 13 (3) (1995) 253–263.
[13] A.J. Fox, Outliers in time series, J. R. Stat. Soc. 34 (B) (1972) 350–363.
[14] P.H. Franses, M. Ooms, C.S. Bos, Long memory and level shifts: re-analyzing inflation rates, Empir. Econom. 24 (3) (1999) 427–449.
[15] J.S. Geweke, Porter-Hudak, The estimation and application of long memory times series model, J. Time Ser. Anal. 4 (4) (1983) 221–238.
[16] L. Giraitis, R. Leipus, A generalized fractionally diferencing approach in long-memory modeling, Lith. Math. J. 35 (1995) 53–65.
[17] C.W.J. Granger, R. Joyeux, An introduction to long-memory times series models and fractional differencing, J. Time Ser. Anal. 1 (1980) 15–29.

[18] U. Hassler, (Mis)specification of long memory in seasonal time, J. Time Series Anal. 15 (1) (1994) 19–30.
[19] J.R. Hosking, Fractional differencing, Biometrika 68 (1981) 165–176.
[20] N.-J. Hsu, H. Tsai, Semiparametric estimation for seasonal long-memory time series using generalized exponential models, J. Statist. Plann. Inference 139 (6) (2009) 1992–2009.
[21] C.M. Hurvich, B.K. Ray, Estimation of the memory parameter for nonstationary or noninvertible fractionally integrated processes, J. Time Ser. Anal. 16 (1) (1995) 17–41.
[22] P. Iglesias, H. Jorqueira, W. Palma, Data analysis using regression model with missing observations and long memory, Comput. Statist. Data Anal. 50 (8) (2006) 2028–2043.
[23] P.S. Kanaroglou, M.D. Adams, P.F.D. Luca, D. Corr, N. Sohel, Estimation of sulfur dioxide air pollution concentrations with a spatial autoregressive model, Atmos. Environ. 79 (2013) 421–427.
[24] C. Lévy-Leduc, H. Boistard, E. Moulines, M.S. Taqqu, V.A. Reisen, Asymptotic properties of U-processes under long-range dependence, Ann. Statist. 39 (3) (2011) 1399–1426.
[25] C. Lévy-Leduc, H. Boistard, E. Moulines, M.S. Taqqu, V.A. Reisen, Large sample behaviour of some well-known robust estimators under long-range dependence, Statistics 45 (1) (2011) 59–71.
[26] C. Lévy-Leduc, H. Boistard, E. Moulines, M.S. Taqqu, V.A. Reisen, Robust estimation of the scale and the autocovariance function of Gaussian short and long-range dependent processes, J. Time Ser. Anal. 32 (2) (2011) 135–156.
[27] I. Lobato, P.M. Robinson, Averaged periodogram estimation of long memory, J. Econometrics 73 (1996) 303–324.
[28] M. Luvsan, R. Shie, T. Purevdorj, L. Badarch, B. Baldorj, C. Chan, The influence of emission sources and meteorological conditions on SO_2 pollution in Mongolia, Atmos. Environ. 61 (2012) 542–549.
[29] Y. Ma, M.G. Genton, Highly robust estimation of the autocovariance function, J. Time Ser. Anal. 21 (6) (2000) 663–684.
[30] G. Marques, Empirical aspects of the Whittle-based maximum likelihood method in jointly estimating seasonal and non-seasonal fractional integration parameters, Physica A 390 (1) (2011) 8–17.
[31] F.F. Molinares, V.A. Reisen, F. Cribari-Neto, Robust estimation in long-memory processes under additive outliers, J. Statist. Plann. Inference 139 (8) (2009) 2511–2525.
[32] B.P. Olbermann, R.C.S. Lopes, V.A. Reisen, Invariance of the first difference in ARFIMA models, Comput. Stat. 21 (3) (2006) 445–461.
[33] W. Palma, Long-Memory Time Series: Theory and Methods, Wiley, 2007.
[34] W. Palma, N. Chan, Efficient estimation of seasonal long-range-dependent processes, J. Time Ser. Anal. 2 (6) (2005) 863–892.
[35] S. Porter-Hudak, An application of the seasonal fractionally differenced model to the monetary aggregates, J. Amer. Statist. Assoc. 85 (410) (1990) 338–344.
[36] M.B. Priestley, Spectral Analysis and Time Series, Academic Press, London, 1981.
[37] R Development Core Team, R: a language and enviroment for statistical computing, Vienna, Austria, 2014.
[38] B. Ray, Long-range forecasting of IBM product revenues using a seasonal fractionally differenced ARMA model, Int. J. Forecast. 9 (2) (1993) 255–269.
[39] V.A. Reisen, Estimation of the fractional difference parameter in the Arima(p, d, q) model using the smoothed periodogram, J. Time Ser. Anal. 15 (1994) 335–350.
[40] V.A. Reisen, C. Lévy-Leduc, M.S. Taqqu, An M-estimator for the long-memory parameter, J. Statist. Plann. Inference 187 (2017) 44–55.
[41] V.A. Reisen, S. Lopes, Some simulations and applications of forecasting long-memory time-series models, J. Statist. Plann. Inference 80 (1999) 269–287.
[42] V.A. Reisen, A. Rodrigues, W. Palma, Estimation of seasonal fractionally integrated processes, Comput. Statist. Data Anal. 50 (2006) 568–582.
[43] V.A. Reisen, A. Rodrigues, W. Palma, Estimating seasonal long-memory processes: A Monte Carlo study, J. Stat. Comput. Simul. 76 (4) (2006a) 305–316.
[44] V.A. Reisen, A.J.Q. Sarnaglia, N.C. Reis Jr, C. Lévy-Leduc, J.M. Santos, Modeling and forecasting daily average PM_{10} concentrations by a seasonal long-memory model with volatility, Environ. Model. Softw. 51 (2014) 286–295.
[45] V.A. Reisen, B. Zamprogno, W. Palma, J. Arteche, A semiparametric approach to estimate two seasonal fractional parameters in the Sarfima model, Math. Comput. Simulation 98 (2014) 1–17.
[46] P.J. Rousseeuw, C. Croux, Explicit scale estimators with high breakdown point, in: L1 Stat. Anal. Relat. Methods, 1992, pp. 77–92.
[47] P.J. Rousseeuw, C. Croux, Alternatives to the median absolute deviation, J. Amer. Statist. Assoc. 88 (424) (1993) 1273–1283.
[48] A. Saral, F. Ertürk, Prediction of ground level SO_2 concentration using artificial neural networks, Water Air Soil Pollut. 3 (5–6) (2003) 307–316.
[49] A. Sarnaglia, V.A. Reisen, C. Lévy-Leduc, Robust estimation of periodic autoregressive processes in the presence of additives outliers, J. Multivariate Anal. 2 (2010) 2168–2183.
[50] M.R. Sena Jr., V.A. Reisen, S.R.C. Lopes, Correlated error in the parameters estimation of the ARFIMA model: a simulated study, Commun. Stat. Simul. Comput. 35 (2006) 789–802.
[51] M.S. Taqqu, Fractional brownian motion and long-range dependence, in: P. Doukhan, G. Oppenheim, M. Taqqu (Eds.), Theory and Applications of Long-Range Dependence, Birkhauser, Boston, 2003, p. 720.
[52] J. Tolvi, Long memory and outliers in stock market returns, Appl. Financ. Econ. 13 (7) (2003) 495–502.
[53] H. Tsai, H. Rachinger, E.M.H. Lin, Inference of seasonal long-memory time series with measurement error, Scand. J. Stat. 42 (1) (2015) 137–154.
[54] C. Velasco, Non-Gaussian log-periodogram regression, Econom. Theory 16 (2000) 44–79.
[55] X. Ye, P. Gao, H. Li, Improving estimation of the fractionally differencing parameter in the Sarfima model using tapered periodogram, Econ. Modell. 46 (2015) 167–179.