Frontiers

Identification and validation of stable ARFIMA processes with application to UMTS data

Krzysztof Burnecki*, Grzegorz Sikora

Faculty of Pure and Applied Mathematics, Hugo Steinhaus Center, Wroclaw University of Science and Technology, Wyspianskiego 27, 50-370 Wroclaw, Poland

A B S T R A C T

In this paper we present an identification and validation scheme for stable autoregressive fractionally integrated moving average (ARFIMA) time series. The identification part relies on a recently introduced estimator which is a generalization of that of Kokoszka and Taqqu and a new fractional differencing algorithm. It also incorporates a low-variance estimator for the memory parameter based on the sample mean-squared displacement. The validation part includes standard noise diagnostics and backtesting procedure. The scheme is illustrated on Universal Mobile Telecommunications System (UMTS) data collected in an urban area. We show that the stochastic component of the data can be modeled by the long memory ARFIMA. This can help to monitor possible hazards related to the electromagnetic radiation.

1. Introduction

The concept of anomalous diffusion and fractional dynamics has deeply penetrated the statistical and chemical physics communities, yet the subject has also become a major field in mathematics [1,2]. Historically, fractional dynamical systems are related to the concept of fractional dynamic equations. This is an active field of study in physics, mechanics, mathematics, and economics investigating the behavior of objects and systems that are described by using differentiation of fractional orders. The celebrated fractional Fokker–Planck equation (FFPE), describing anomalous diffusion in the presence of an external potential was derived explicitly in [3], where methods of its solution were introduced and for some special cases exact solutions were calculated.

Derivatives and integrals of fractional orders can be used to describe random phenomena that can be characterized by long (power-like) memory or self-similarity [1,2]. Long memory (or long-range dependence) is a property of certain stationary stochastic processes describing phenomena, which concern the events that are arbitrarily distant still influence each other exceptionally strong. It has been associated historically with slow decay of correlations and a certain type of scaling that is connected to self-similar processes [4,5].

Recently, there has been a great interest in long-range dependent and self-similar processes, in particular fractional Brow-

nian motion (FBM), fractional stable motion (FSM) and autoregressive fractionally integrated moving average (ARFIMA), which are also called fractional autoregressive integrated moving average (FARIMA) [6,7]. This importance can be judged, for example, by a very large number of publications having one of these notions in the title, in areas such as finance and insurance [8–15], telecommunication [16–21], hydrology [22], climate studies [23], linguistics [24], DNA sequencing [25] or medicine [26]. Long-range dependent and self-similar processes also appear widely in other areas like biophysics [7,27–32] or astronomy [33]. These publications address a great variety of issues: detection of long memory and self-similarity in the data, statistical estimation of parameters of long-range dependence and self-similarity, limit theorems under long-range dependence and self-similarity, simulation of long memory and self-similar processes, relations to ergodicity and many others [6,7,34–37].

The FBM, FSM and ARFIMA serve as basic stochastic models for fractional anomalous dynamics [7]. The former two models are self-similar and their increments form long-range dependent processes. The discrete-time ARFIMA process is stationary and generalizes both models since aggregated, in the limit, it converges to either fractional Brownian or stable motion. As a consequence, a partial sum ARFIMA process can be considered as a unified model for fractional anomalous diffusion in experimental data [38]. A type of anomaly of the process is controlled only by its the memory parameter regardless of the underlying distribution [28]. We also note that there is a relationship between the ARFIMA and continuous time random walk (CTRW) which is a classical model of anomalous diffusion [3,39]. The latter can be obtained by subordi-

nation of the Ornstein–Uhlenbeck process which discrete version is an autoregressive (AR) process, so a special case of the ARFIMA [40,41].

In contrast to FBM and FLSM, ARFIMA allows for different light- and heavy-tailed distributions, and both long (power-like) and short (exponential) dependencies [38]. Moreover, as a stationary process, it provides prediction tools.

It appears that the values of ARFIMA with Gaussian noise, for the memory parameter d greater than 0, have so slowly decaying autocovariance function that it is not absolutely summable. This behavior serves as a classical definition of the long-range dependence. However, it is also a well-known fact that the heavy-tailed probability distributions with diverging variance are ubiquitous in nature and finance [42–47].

The stable probability densities have the asymptotics decaying at infinity as $|x|^{-1-\alpha}$, where α is the index of stability varying between 0 and 2. They attract distributions having the same law of decay. On the contrary, the Gaussian distribution has the index of stability 2 and attracts all distributions with lighter tails [42,48,49].

Stably distributed random noises are observed in such diverse applications as plasma turbulence (density and electric field fluctuations [49–51]), stochastic climate dynamics [52–54], physiology (heartbeats [55]), electrical engineering [56], biology [28,30], and economics [57,58]. Heavy-tailed distributions govern circulation of dollar bills [59] and behavior of the marine vertebrates in response to patchy distribution of food resources [60].

In this paper we propose an identification and validation scheme for ARFIMA processes with noise in the domain of attraction of the stable law which is based on estimation algorithm introduced in [61]. The scheme is illustrated on the electromagnetic radiation data which shows long memory behavior which is also observed for telecommunication data in [19].

The paper is organized as follows: in Section 2 we recall basic facts about a prominent example of long memory processes, namely ARFIMA time series. In Section 3 we introduce a step by step procedure for identification of a ARFIMA process. The procedure involves (i) a method of preliminary estimation of the memory parameter based on the mean-squared displacement, (ii) a new method of fractional differencing which leads to model order estimation and (iii) the estimation formula for stable ARFIMA times series introduced in [61]. Section 4 is devoted to validation of the fitted model. It consists of analysis of residuals: testing their randomness and fitting a distribution which is done by standard statistical tests, and backtesting which involves prediction formula for ARFIMA time series. The identification and validation procedure is illustrated in Section 5 on electromagnetic field data collected in the vicinity of an Universal Mobile Telecommunications System (UMTS) station in Wroclaw. After removing deterministic seasonality and volatility from the data, a long memory ARFIMA process is identified and validated. In Section 6 a summary of the results is given.

2. ARFIMA process

In this section we briefly present the main facts about ARFIMA time series which were introduced in [62] and [63]. Such process $\{X_t\}$, denoted by ARFIMA(p, d, q), is defined by

$$\Phi_p(B)X_t = \Theta_q(B)(1-B)^{-d}Z_t, \tag{1}$$

where innovations (noise sequence) Z_t are i.i.d. random variables with either finite or infinite variance. We also assume that the innovations belong to the domain of attraction of an α-stable law with $0 < \alpha \le 2$. For the infinite variance case ($\alpha < 2$) this means that

$$P(|Z_t| > x) = x^{-\alpha}L(x), \ as \ x \to \infty, \tag{2}$$

where L is a slowly varying function at infinity, and

$$\frac{P(Z_t > x)}{P(|Z_t| > x)} \to a, \quad \frac{P(Z_t < -x)}{P(|Z_t| > x)} \to b, \ as \ x \to \infty, \tag{3}$$

where a and b are nonnegative numbers such that $a + b = 1$. The finite variance case ($\alpha = 2$) leads the domain of attraction of Gaussian law. Polynomials Φ_p and Θ_q have classical forms, i.e. $\Phi_p(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p$ is the autoregressive polynomial, $\Theta_q(z) = 1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q$ is the moving average polynomial. The operator B, called the backward operator, satisfies $BX_t = X_{t-1}$ and $B^j X_t = X_{t-j}, \ j \in \mathbb{N}$. The crucial part of ARFIMA Definition (1) is the operator $(1-B)^{-d}$ called the fractional integrating operator and the fractional number d called the memory parameter.

The operator $(1-B)^{-d}$ has the infinite binomial expansion

$$(1-B)^{-d}Z_t = \sum_{j=0}^{\infty} b_j(d)Z_{t-j}, \tag{4}$$

where the $b_j(d)'$s are the coefficients in the expansion of the function $f(z) = (1-z)^{-d}, \ |z| < 1$, i.e.

$$b_j(d) = \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)}, \ j = 0, 1, \ldots, \tag{5}$$

where Γ is the gamma function. The series (4) is convergent almost surely and ARFIMA Definition (1) is correct if and only if

$$\alpha(d-1) < -1 \Leftrightarrow d < 1 - \frac{1}{\alpha}. \tag{6}$$

In the Gaussian case where $\alpha = 2$ we have $d < 1/2$. Under Assumption (6) and when polynomials Φ_p and Θ_q do not have common roots, and Φ_p has no roots in the closed unit disk $\{z: |z| \le 1\}$, the ARFIMA(p, d, q) time series defined by (1) has the causal moving average form

$$\underbrace{X_t}_{\text{ARFIMA } (p,d,q)} = \overbrace{(1-B)^{-d}}^{d-\text{fractional integrating}} \underbrace{\frac{\Theta_q(B)}{\Phi_p(B)}}_{\text{ARMA } (p,q)} Z_t = \sum_{j=0}^{\infty} c_j(d)Z_{t-j}, \tag{7}$$

where

$$c_j(d) = b_j(d) + \sum_{i=1}^{p} \phi_i c_{j-i}(d) + \sum_{k=1}^{q} \theta_k b_{j-k}(d), \tag{8}$$

for $j = 0, 1, \ldots$ Therefore ARFIMA(p, d, q) time series can be obtained by d-fractional integrating of ARMA(p, q) series. The d-fractional integrating through $(1-B)^{-d}$ operator builds the dependence between observations in a ARFIMA sequence, even as they are far apart in time.

When $d > -1 + 1/\alpha$ the ARFIMA(p, d, q) time series $\{X_t\}$ is invertible and Definition (1) can be rewritten in the equivalent form

$$\Phi_p(B) \underbrace{\overbrace{(1-B)^d}^{d-\text{frac. differencing}} \overbrace{X_t}^{\text{ARFIMA } (p,d,q)}}_{\text{ARMA } (p,q)} = \Theta_q(B)Z_t. \tag{9}$$

The operator $(1-B)^d$, called the fractional differencing operator, is the inverse operator of the fractional integrating operator $(1 - B)^{-d}$. It has the infinite binomial expansion of the form (4) with the opposite d. Hence, according to (9), the ARMA(p, q) time series can be obtained after d–fractional differencing of ARFIMA(p, d, q) sequence.

When the memory parameter d is close to 1/2, all the coefficients $b_j(d)'$s are positive and converge to zero at a power rate. In view of series representation (4), ARFIMA(0, d, 0) observation X_t depends not only on the present noise observation Z_t, but also depends strongly on the whole history of the noise process. Hence

in ARFIMA(0, d, 0) series with $d \approx 1/2$ observations that are far apart in time still depend significantly from each other. This phenomenon is called in the literature the long memory property, see [6], and for the Gaussian time series, i.e. $\alpha = 2$, its autocovariance function Cov(t) satisfies

$$\sum_{t=0}^{\infty} |Cov(t)| = \infty. \tag{10}$$

For the Gaussian ARFIMA the rate of decay of the covariance function is t^{2d-1} (see [63]), so for $d > 0$ long memory property (10) is satisfied. For processes with $\alpha < 2$ the covariance is replaced by the codifference

$$\tau(X_t, X_0) = \ln Ee^{(i(X_t - X_0))} - \ln Ee^{iX_t} - \ln Ee^{-iX_0},$$

which reduces to covariance in the Gaussian case, for details see [64]. For stable ARFIMA series the rate of decay of the codifference function is $t^{\alpha(d-1)+1}$ and therefore the long memory property (10) holds for $d > 1 - 2/\alpha$, for details see [65].

3. ARFIMA identification

In this section we describe the identification algorithm of ARFIMA processes. In this procedure we assume that data that are stationary.

3.1. Memory parameter estimation

The first step in the ARFIMA identification is to investigate the dependence structure of our sample data. In ARFIMA models the strength of dependence is reflected by the memory parameter d. We now present three methods of estimation of the parameter d which will be used in further analyzes: two classic and the one which was recently introduced in the literature. We also note that there are other methods of estimating the memory parameter, like, e.g., the GPH estimator [66]. In general, we advocate using a few of them and comparing their results.

3.1.1. The RS method
It is a classical method of investigating the long memory and it was developed by Hurst in [67]. For α-stable time series (including Gaussian case) the method returns $d + 1/2$, for details see [68]. The algorithm starts with dividing the data set $\{X_i : i = 1, 2, \ldots, N\}$ of length N into blocks of equal length n. Then, for each subseries $(m = 1, 2, \ldots, [N/n])$ the $(R/S)_m$ statistics is computed according to the formula

$$(R/S)_m = \frac{\max \sum_{i=1}^{k} (X_i - \overline{X}_m) - \min \sum_{i=1}^{k} (X_i - \overline{X}_m)}{S_m}, \tag{11}$$

where max and min are taken for $1 \leq k \leq n$, $\overline{X}_m$ and S_m are sample mean and sample standard deviation of the mth subseries respectively. Next, the sample mean of the $(R/S)_m$ statistics over blocks is calculated and denoted by $(R/S)_n$. Finally, the $(R/S)_n$ statistics are plotted against n on a double–logarithmic scale with the fitted least squares line. The slope should be equal to $d + 1/2$.

3.1.2. Lo's modified RS method
The Lo's modified R/S statistics (MRS), proposed in [12], is defined by

$$V_k = \sqrt{N}(R/S_k)_N,$$

which is a modification of the formula (11) for $n = N$ with the denominator

$$S_k = \sqrt{S^2 + 2 \sum_{i=1}^{k} w_i(k) \hat{\gamma}_i},$$

where $\hat{\gamma}_i$ are the sample autocovariances and weights $w_i(k)$ are given by

$$w_i(k) = 1 - \frac{i}{k+1}, \quad k < N.$$

Plotting V_q with respect to q on a log-log scale and fitting the least squares line, leads, for the finite variance case, to an estimate of H, namely the slope of the line equals $1/2 - H$. One may check that, as it was the case for R/S statistic, that the formula for the α-stable distribution can be generalized to $-d$, where $d = H - 1/\alpha$. For details see [69].

3.1.3. Sample mean-squared displacement
For the finite sample $\{X_i : i = 1, 2, \ldots, N\}$ with stationary increments the sample mean-squared displacement (sample MSD) has the following formula

$$M_N(\tau) = \frac{1}{N-\tau} \sum_{k=1}^{N-\tau} (X_{k+\tau} - X_k)^2.$$

The sample MSD is a time average MSD on a finite sample regarded as a function of difference τ between observations. It was shown in [28] that if the sample comes from an H-self-similar α-stable process with stationary increments then for large N

$$M_N(\tau) \stackrel{d}{\sim} \tau^{2d+1}, \tag{12}$$

where $d = H - 1/\alpha$ and $\stackrel{d}{\sim}$ means similarity in distribution. The sample MSD can serve as a method of estimating d. Applying (12), one can fit the linear regression line on a log-log scale to the values of sample MSD. The slope of the fitted line equals $2d + 1$.

3.2. Fractional differencing – from ARFIMA to ARMA

Knowing already the strength of dependence in our sample data, i.e. the estimated value of the memory parameter d, we can get rid of such dependence by fractional differencing. We briefly describe here the issue of fractional differencing, especially in practical aspects. Recall that since (9), one can d–fractionally differentiate the ARFIMA(p, d, q) sequence $\{X_t\}$ and obtain the ARMA(p, q) series $\{(1 - B)^d X_t\}$. The motivation for such operation is the fact that ARMA models are easier objects to study.

Our idea of fractional differencing is to apply the simulation algorithm of ARFIMA series, introduced in [34]. The method relies on properties of Discrete Fourier Transforms (DFT). For M-periodic sequence $a = \{a_j : j \in \mathbb{Z}\}$, i.e. $a_{j+M} = a_j$, for all $j \in \mathbb{Z}$, the DFT $\mathcal{D}_M(a) = \{\mathcal{D}_M(a)_k : k \in \mathbb{Z}\}$ is defined as follows

$$\mathcal{D}_M(a)_k := \sum_{j=0}^{M-1} e^{2\pi ijk/M} a(j), \quad k \in \mathbb{Z},$$

with the inverse formula

$$a_j = \mathcal{D}_M^{-1}(\mathcal{D}_M(a))_j := \frac{1}{M} \sum_{k=0}^{M-1} e^{-2\pi ijk/M} \mathcal{D}_M(a)_k, \quad j \in \mathbb{Z}.$$

Moreover, for any two M–periodic sequences a and b, the following convolution theorem holds

$$\mathcal{D}_M(a)_k \mathcal{D}_M(b)_k = \mathcal{D}_M(a * b)_k, \quad k \in \mathbb{Z}, \tag{13}$$

where

$$(a * b)_n := \sum_{j=0}^{M-1} a_{n-j} b_j, \quad n \in \mathbb{Z}.$$

If M is an integer power of two, then to compute the DFT of M–periodic sequence one can apply efficient FFT (radix–2) algorithm,

for details see [70]. The proposed fractional differencing algorithm is based on the approximation of the ARMA path of the length N:

$$Y_n := (1 - B)^d X_n = \sum_{j=0}^{\infty} b_j(-d)Z_{n-j}, \ n = 0, 1, \ldots, N - 1,$$

by the truncated moving average

$$Y_n \approx Y_n^M := \sum_{j=0}^{M-1} b_j(-d)Z_{n-j}, \ n = 0, 1, \ldots, N - 1. \quad (14)$$

Because R.H.S. of formula (14) has a form like finite discrete convolution, the main idea is to apply convolution theorem (13) and efficient FFT techniques. First we define $(M + N - 1)$-periodic sequence $\tilde{b}_{j+k(M+N-1)}(-d) := \tilde{b}_j(-d), \ j = 0, 1, \ldots, M + N - 2, \ k \in \mathbb{Z}$, where

$$\tilde{b}_j(-d) = \begin{cases} b_j(-d), & \text{for } j = 0, 1, \ldots, M - 1, \\ 0, & \text{for } j = M, M + 1, \ldots, M + N - 2, \end{cases}$$

with $b_j(-d)$'s computed according to formula (5). To apply convolution techniques as in the simulation algorithm, for a ARFIMA trajectory $\{X_n : n = 0, 1, \ldots, M + N - 2\}$ of length $M + N - 2$ we define $(M + N - 1)$-periodic sequence $\tilde{X}_{j+k(M+N-1)} := X_j, \ j = 0, 1, \ldots, M + N - 2, \ k \in \mathbb{Z}$. Therefore, Y_n can be approximated by the convolution of these two periodic sequences:

$$Y_n \approx \sum_{j=0}^{M+N-2} \tilde{b}_j(-d)\tilde{X}_{n-j}, \ n = 0, 1, \ldots, N - 1$$

and computed by taking the inverse DFT of the product of the DFT's of the two $(M + N - 1)$-periodic sequences $\tilde{b} = \{\tilde{b}_j(-d) : j \in \mathbb{Z}\}$ and $\tilde{X} = \{\tilde{X}_j : j \in \mathbb{Z}\}$. The whole algorithm reads as follows.

Fractional differencing algorithm

Step 1. Choose a truncation parameter $M \in \mathbb{N}$ and expected sample size $N \in \mathbb{N}$, such that $M + N - 1$ is the length of data set and is an integer power of two.

Step 2. According to relation (5), compute $(M + N - 1)$-periodic sequence $\tilde{b} = \{\tilde{b}_j(-d) : j = 0, 1, \ldots, M + N - 2\}$ and using the radix–2 FFT algorithm, compute the DFT $\mathcal{D}_{M+N-1}(\tilde{b})$ of the sequence $\tilde{b}$.

Step 3. Using the radix–2 FFT algorithm, compute the DFT $\mathcal{D}_{M+N-1}(\tilde{X})$ of the $(M + N - 1)$–periodic sequence $\tilde{X}$.

Step 4. By using the radix–2 FFT algorithm, compute the inverse DFT of the product $\mathcal{D}_{M+N-1}(\tilde{b})\mathcal{D}_{M+N-1}(\tilde{X})$ and approximate Y_n by

$$Y_n \approx \mathcal{D}_{M+N-1}^{-1}(\mathcal{D}_{M+N-1}(\tilde{b})\mathcal{D}_{M+N-1}(\tilde{X}))_n,$$

for $n = 0, 1, \ldots, N - 1$, where the sequence product $\mathcal{D}_{M+N-1}(\tilde{b})\mathcal{D}_{M+N-1}(\tilde{X})$ consists of numbers

$$\mathcal{D}_{M+N-1}(\tilde{b})_j \mathcal{D}_{M+N-1}(\tilde{X})_j,$$

$$j = 0, 1, \ldots, M + N - 2.$$

In the literature different methods of fractional differencing have been proposed, see, e.g. [71–73]. It may seem that our algorithm has a major drawback from a practical point of view. Note that in order to obtain a series of length N we have to fractionally differentiate a sequence of length $N + M - 1$. Therefore, we loose a large part of the data (especially when $M \gg N$). However, we believe that it is necessary to truly differentiate data in a fractional way. This can be explained intuitively. Since the fractional differentiation leads to getting rid of the dependence between distant observations, we can remove such dependence from late observations by removing the effect of initial observations. However, we can not fractionally differentiate initial data because we do not have earlier history.

3.3. ARMA order criteria

The fractional differencing operation leads from ARFIMA to ARMA series, recall Eq. (9). Hence, our sample data to which we try to fit the ARFIMA process, now fractionally differentiated, can be treated as an ARMA sequence.

The first task in fitting the ARMA model to data is to determine the order of such model, i.e. to find degrees p and q of the polynomials Φ_p and Θ_q respectively. The most common criterion is the bias–corrected version of the Akaike criterion, i.e. (p, q) should be selected to minimize the AICC statistics

$$AICC(p, q) := -2\ln L + \frac{2(p + q + 1)N}{N - p - q - 2}, \quad (15)$$

where L denotes the maximum likelihood function and N the length of the sample. For details of the maximum likelihood estimation and order criteria see [75]. One can also additionally use visual techniques based on autocorrelation function (ACF) and partial autocorrelation function (PACF).

3.4. ARFIMA parameter estimation

We describe here an estimation method for parameters of ARFIMA(p, d, q) model in the situation when the order (p, q) is known. This method, based on a variant of Whittle's estimator [35,74], was developed in [76] for ARFIMA time series with innovations satisfying conditions (2) and (3) and with long memory, i.e. the memory parameter $d \in (0, 1/2)$ and $\alpha \in (1, 2]$. In [61], this algorithm was generalized to the case of short memory, i.e. the memory parameter $d \in (-1/2, 0)$ and $\alpha \in (2/3, 1]$. Thus, this estimation methodology can be applied to both short and long memory ARFIMA processes.

We want to estimate the $(p + q + 1)$–dimensional vector

$$\beta_0 := (\phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q, d),$$

where $\phi_1, \phi_2, \ldots, \phi_p$ and $\theta_1, \theta_2, \ldots, \theta_q$ are the coefficients of the polynomials Φ_p and Θ_q respectively. The vector β_0 is from the parameter space

$$E := \Big\{ \beta : \phi_p, \theta_q \neq 0, \Phi_p(z)\Theta_q(z) \neq 0 \text{ for } |z| \leq 1,$$

$$\Phi_p, \Theta_q \text{ have no common roots}, d \in \left(0, 1 - \frac{1}{\alpha}\right) \Big\}.$$

We introduce the normalized periodogram

$$\tilde{I}_n(\lambda) := \frac{\left|\sum_{t=1}^{n} X_t e^{-i\lambda n}\right|^2}{\sum_{t=1}^{n} X^2(t)}, \ -\pi \leq \lambda \leq \pi, \quad (16)$$

and the quantity

$$\gamma(0) := \sum_{j=0}^{\infty} c_j^2(d), \quad (17)$$

where the coefficients $c_j(d)$'s are defined in (7). We also define the power transfer function

$$g(\lambda, \beta) := \left| \frac{\Theta_q(e^{-i\lambda}, \beta)}{\Phi_p(e^{-i\lambda}, \beta)(1 - e^{-i\lambda})^{d(\beta)}} \right|^2.$$

The estimator β_n of the true parameter vector β_0 is defined as

$$\beta_n := \arg\min_{\beta \in E} \int_{-\pi}^{\pi} \frac{\tilde{I}_n(\lambda)}{g(\lambda, \beta)} d\lambda. \quad (18)$$

The method is justified by the consistency result, see [76] in the long memory case and [61] in the short memory case.

The presented estimation algorithm can be applied twice: to ARFIMA trajectory and to fractionally differentiated ARFIMA trajectory (which is simply ARMA trajectory). The estimated parameters

should be very close, except the memory parameter d, which is close to zero for fractionally differentiated ARFIMA sequences. We illustrate such double application of this method in the next section for the UMTS data. Finally, the estimated d for a ARFIMA trajectory should be close to the previously obtained values of the memory parameter d in Section 3.1.

4. ARFIMA validation

In this section we describe statistical tools that can be applied to justify the hypothesis of ARFIMA time series as an underlying model for empirical data.

4.1. Tests of randomness of residuals

Having the results of the parameter estimation, in order to justify the ARFIMA model we calculate the residuals and check their independence. Recall that according to (9), one can get the noise sequence:

$$Z_n = \frac{\Phi_p(B)}{\Theta_q(B)}(1 - B)^d X_n, \quad n = 1, 2, \ldots, N.$$

Here, we present statistical tests for checking the randomness of the residual sequence.

4.1.1. Ljung-Box test

Ljung and Box introduced in [77] the statistical test which takes into account magnitude of correlation as a group. The test statistics is

$$Q = N(N + 2) \sum_{i=1}^{K} \frac{\hat{\rho}_i}{N - i},$$

where $\hat{\rho}_i$ is a squared sample autocorrelation at lag i and K is the number of lags included in the statistic. Under the null hypothesis of i.i.d. sequence the Q statistic is approximately distributed as $\chi^2(K)$ with K degrees of freedom. Therefore large values of Q statistic indicate that the sample autocorrelation of data is too large for i.i.d. sequence. Hence, one reject the null hypothesis at confidence level α if $Q > \hat{\chi}^2_{1-\alpha}(K)$, where $\hat{\chi}^2_{1-\alpha}(K)$ is the $(1 - \alpha)$ quantile from the $\chi^2(K)$ distribution. For different K values one may testing the strength of dependence in the corresponding group of observations.

4.1.2. Turning points test

For the finite sample $\{X_i : i = 1, 2, \ldots, N\}$, we say that there is a turning point at time i $(1 < i < n)$ if $X_{i-1} < X_i$ and $X_i > X_{i+1}$ or if $X_{i-1} > X_i$ and $X_i < X_{i+1}$. Under the null i.i.d. hypothesis the turning points number T is approximately $\mathcal{N}(\mu_T, \sigma_T)$, where $\mu_T = 2(n - 2)/3$ and $\sigma_T^2 = (16n - 29)/90$. One reject the null hypothesis at level α if $|T - \mu_T|/\sigma_T > \Psi_{1-\alpha/2}$, where $\Psi_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution. For details see [75].

4.2. Residual distribution fitting

The next challenge with the estimated noise sequence, following checking the randomness, is the distribution fitting. We present here briefly various statistical tests based on empirical distribution function. In those testing procedures different statistics measure the vertical distance between the empirical $F_N(x)$ and theoretical $F(x)$ cumulative distribution function. The obtained values of such statistics are used to compute so-called p-values. They are probabilities of obtaining a value of test statistic at least as extreme as the one that was actually observed for the sample data, assuming that the hypothesis is true. Therefore the lower the p-value, the less likely the result is if the hypothesis of underlying distribution

is true. Generally one rejects the hypothesis about the distribution when the p-value is less than 0.05 or 0.01.

In this paper we consider two classes of such tests, namely the Kolomogorov–Smirnov and Cramér–von Mises. The Kolmogorov–Smirnov statistic is given by

$$D = \sup_x \{|F_N(x) - F(x)|\}.$$

It can be written as a maximum of two nonnegative supremum statistics:

$$D^+ = \sup_x \{F_N(x) - F(x)\} \text{ and } D^- = \sup_x \{F(x) - F_N(x)\}.$$

We also use the Kuiper statistics from the same class, given by

$$V = D^+ + D^-.$$

The second class of vertical measures is the Cramér–von Mises family

$$CM = N \int_{-\infty}^{\infty} (F_N(x) - F(x))^2 \psi(x) dF(x).$$

The function $\psi(x)$ is a special weight function. For $\psi(x) = 1$ we obtain the Cramér–von Mises statistic and for $\psi(x) = [F(x)(1 - F(x))]^{-1}$ we arrive at A^2 Anderson and Darling one.

Here we also present the methodology of computing the p-values. The procedure is based on Monte-Carlo simulations [78]. First, from studied empirical sample we estimate a vector of parameters $\hat{\theta}$ of suspected underlying distribution. Then, we calculate a test statistic. Next, we generate a large number. e.g. 1000, samples of the same length as empirical data from suspected distribution with estimated parameters $\hat{\theta}$ and for each calculate the value of test statistic. Finally, the p-value is a proportion of times that the statistic value is at least as for original sample data.

In this paper we consider four possible probability laws underlying the data: Gaussian, α-stable, normal-inverse Gaussian (NIG) and tempered α-stable. The random variable X has an α-stable distribution if there are parameters $\alpha \in (0, 2]$, $\sigma \leq 0$, $\beta \in [-1, 1]$ and real μ such that its characteristic function has the form:

$$E\exp\{i\theta X\} = \begin{cases} \exp\left\{-\sigma^\alpha |\theta|^\alpha \left(1 - i\beta \mathrm{sgn}(\theta) \tan \frac{\pi\alpha}{2}\right) + i\mu\theta\right\}, \\ \exp\left\{-\sigma |\theta| \left(1 + i\beta \frac{2}{\pi} \mathrm{sgn}(\theta) \ln |\theta|\right) + i\mu\theta\right\}, \end{cases}$$

for $\alpha \neq 1$ and $\alpha = 1$ respectively. We denote such distribution by $X \sim S_\alpha(\sigma, \beta, \mu)$. For $\alpha = 2$ we obtain the Gaussian distribution. One of the important properties of such distribution for $\alpha < 2$ are heavy-tails which decay hiperbolically as $x^{-\alpha}$. As a consequence $E|X|^p = \infty$ for any $p \geq \alpha$. The parameter β controls the skewness and the interpretations of σ and μ are similar as for the normal distribution. For other properties of the α-stable distribution we refer to [64].

The NIG distribution, introduced in [79], is chosen as it is able to model both symmetric and asymmetric data with possibly long tails in both directions. Its tail behavior is often classified as "semi-heavy", i.e. the tails are lighter than those of α-stable laws, but much heavier than the Gaussian. Moreover, empirical experience shows an excellent fit of the NIG law to different data [80].

The last considered class of distributions is the family of tempered α-stable distributions. The complete description of this class is given in [81,82]. We consider here one of the members of the tempered stable family, namely the skew tempered α-stable distribution. The random variable T with such distribution is characterized by its characteristic function

$$E\{\exp i\theta T\} = \exp\{(\lambda - i\theta)^\alpha - \lambda^\alpha + i\alpha\lambda^{\alpha-1}\theta\}$$

for $\alpha \neq 1$ and

$$E\{\exp i\theta T\} = \exp\{(\lambda - i\theta)\log(\lambda - i\theta) \\ - \lambda \log(\lambda) + i\theta(1 + \log(\lambda))\}$$
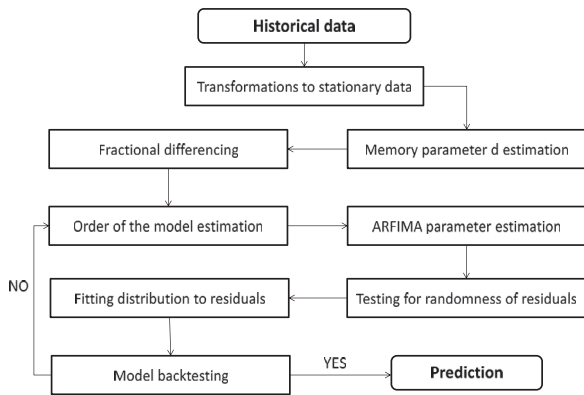
Fig. 1. Identification and validation flow chart for the ARFIMA process.

Fig. 2. UMTS data $\{X_n : n = 1, 2, \ldots, 9999\}$ collected in Wroclaw in 2011.

Fig. 3. Periodogram of UMTS data. The half-day (714) and day (1428) frequencies are clearly visible.

Table 1
Fitted functions $S_k(x)$ with their errors $ERR(S_k)$.

Function $S_k(x)$	Periods $d_1, \ldots, d_k$	$ERR(S_k)$
S_1	$d_1 = 714$	0.1540
	$d_1 = 1428$	0.0933
	$d_1 = 3332$	0.1545
S_2	$\mathbf{d_1 = 714, \ d_2 = 1428}$	**0.0888**
	$d_1 = 1428, \ d_2 = 3332$	0.0891
	$d_1 = 714, \ d_2 = 3332$	0.1544
S_3	$d_1 = 714, \ d_2 = 1428, \ d_3 = 3332$	0.1058

for $\alpha = 1$. The parameter $\lambda > 0$ is called the tempering index. The following connection between probability density functions (PDFs) of tempered and classical stable distributions holds:

$$p_T(x) = \begin{cases} \exp(-\lambda x + (\alpha - 1)\lambda^\alpha) p_U(x - \alpha \lambda^{\alpha-1}), \\ \exp(-\lambda x + \lambda) p_U(x - (1 + \log \lambda)), \end{cases}$$

for $\alpha \neq 1$ and $\alpha = 1$ respectively, where $p_T(\cdot)$ is the PDF of the tempered stable random variable T and $p_U(\cdot)$ is the PDF of the random variable $U \sim S_\alpha(|\cos \frac{\pi\alpha}{2}|^{1/\alpha}, 1, 0)$ for $\alpha \neq 1$ and $U \sim S_1(\frac{\pi}{2}, 1, 0)$ for $\alpha = 1$. In contrast to the classical stable distribution the tempered stable random variable T has all moments finite. Moreover the right tail of such distribution behaves like $e^{-\lambda x}x^{-\alpha}$.

4.3. ARFIMA backtesting

The final step to validate the identified model is the procedure of backtesting. It consists in assessing the accuracy of fitted model's predictors using existing historical data. The backtesting process starts by selecting a threshold date within a time span covered by the historical data. Then, for the threshold, the historical data is truncated at the threshold, the model is fitted to the truncated data and the forecasts are compared with the original untruncated data.

For the proposed class of ARFIMA models the linear predictor based on the historical data $X_n, \ldots, X_0$ is defined as

$$\hat{X}_{n+k} = \sum_{j=0}^{n} a_j X_{n-j},$$

where

$$a_j = -\sum_{t=0}^{k-1} c_j(d) h_{j+k-t}(d),$$

and $h_j(d)$'s are given by

$$\frac{\Phi_p(z)(1-z)^d}{\Theta_q(z)} = \sum_{j=0}^{\infty} h_j(d) z^j, \quad |z| < 1.$$

The coefficients $c_j(d)$ are defined in formula (8). A detailed information on the ARFIMA prediction is given in [83].

All steps of identification and validation of the ARFIMA process are summarized in Fig. 1.

5. UMTS data

In this section we analyze a set of UMTS data, see Fig. 2. The electromagnetic field intensity was measured in Wroclaw in an urban area every minute from 12.01.2011 22:40 to 19.01.2011 21:18 (9999 observations).
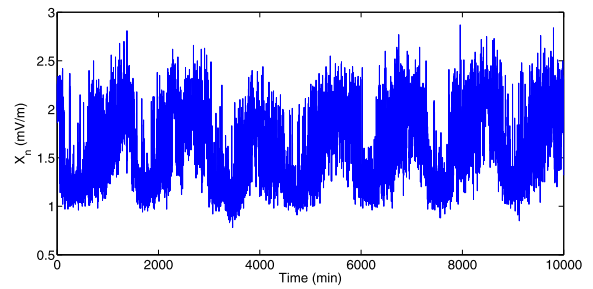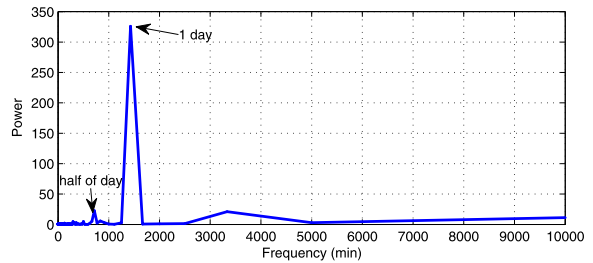
5.1. Seasonality and volatility in UMTS data

We denote the obtained time series by $\{X_n : n = 1, 2, \ldots, 9999\}$. We can clearly notice a strong seasonal component in the data. Therefore, our first task is to remove this seasonality. In order to discover the governing periods we depict the periodogram plot, see Fig. 3. The plot is a sample analogue of the spectral density [75]. In Fig. 3 we can clearly see peaks at frequencies 714, 1428 and 3332. They corresponds to cycles with periods of 12, 24 and 56 h. Since the seasonality structure of data seems to have a sinusoidal form, we try to remove it by fitting a sum of sine functions. The proposed functions have the following form

$$S_k(x) = p_1 + \sum_{i=1}^{k} p_{2i} \sin \left[\frac{2\pi}{d_i} (x + p_{2i+1}) \right],$$

where p_j's are parameters and d_j's are the selected frequencies of the corresponding cycles. We define an error of the fit of the function to the data $\{X_n : n = 1, 2, \ldots, 9999\}$ as

$$ERR(S_k) = \frac{1}{9999} \sum_{n=1}^{9999} (S_k(X_n) - X_n)^2.$$

In Table 1 we present errors related to fitting all possible functions $S_k(x)$ for the three discovered periods: $d_1 = 714$, $d_2 = 1428$ and $d_3 = 3332$. From Table 1 we can see that the lowest error is for the case with sum of two sines having periods 12 and 24 h. Hence, we remove seasonality by subtracting the values of the function $S_2(x)$ with $d_1 = 714$ and $d_2 = 1428$ from the analyzed series, see Fig. 4. We denote the new series by $\{\hat{X}_n : n = 1, 2, \ldots, 9999\}$.
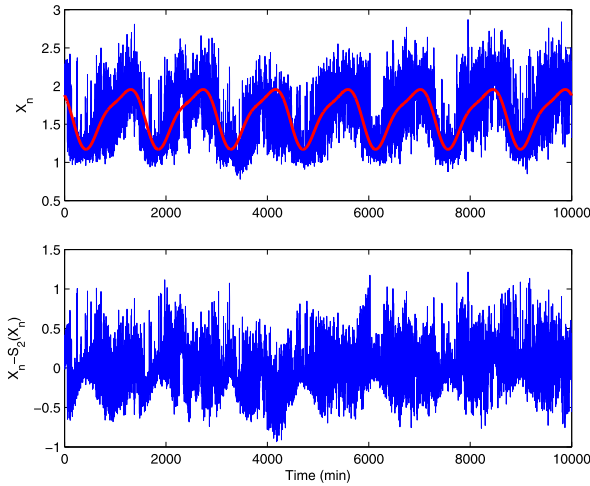
Fig. 4. (Top panel) UMTS data $\{X_n : n = 1, 2, \ldots, 9999\}$ with the fitted sum of sines $S_2(x)$ (red line). (Bottom panel) UMTS data $\{\hat{X}_n : n = 1, 2, \ldots, 9999\}$ after removal of $S_2(x)$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
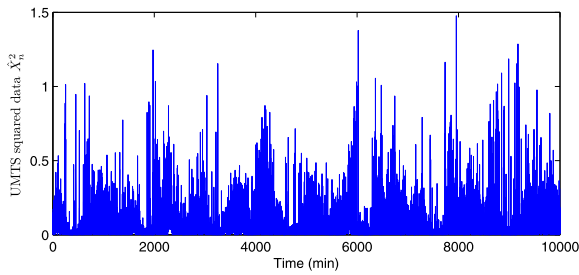

Fig. 5. Squared UMTS data $\{\hat{X}_n^2 : n = 1, 2, \ldots, 9999\}$ with removed seasonality.

After removing seasonality we check stationarity by studying volatility of the data $\{\hat{X}_n : n = 1, 2, \ldots, 9999\}$. A fluctuating variance can be observed in a squared data plot. We present the squared series $\{\hat{X}_n^2 : n = 1, 2, \ldots, 9999\}$ in Fig. 5. We can clearly see that volatility changes in time. In order to remove this effect and make the volatility constant we propose the following procedure based on the moving sample variance. We calculate the sample moving variance corresponding to time intervals of 61 min (61 observations) along the whole series $\{\hat{X}_n^2 : n = 1, 2, \ldots, 9999\}$, i.e.

$$S_k^2 := \frac{1}{60} \sum_{i=k}^{k+60} (\hat{X}_k - \bar{X})^2,$$

for $k = 1, 2, \ldots, 9939$, see Fig. 6. Next, we remove volatility and compute the series $\{\tilde{X}_n := \hat{X}_{n+30}/\sqrt{S_n^2} : n = 1, 2, \ldots, 9939\}$. We notice that the observation $\hat{X}_n$ (for $n = 31, 32, \ldots, 9969$) is divided by the square root of sample variance S_{n-30}^2 from the time interval with 30 observations measured before and after the measurement $\hat{X}_n$. That is why we focus on observations from $n = 31$ to $n = 9969$ and therefore the length of the series $\{\tilde{X}_n : n = 1, 2, \ldots, 9939\}$ is 9939. It seems that such data are closer to constant volatility, see the squared data plot in Fig. 6 (bottom panel). We apply this transformation to the UMTS data and obtain the series $\{\tilde{X}_n = \hat{X}_{n+30}/\sqrt{S_n^2} : n = 1, 2, \ldots, 9939\}$ which is plotted in Fig. 7. We choose it for further studies.

5.2. Modeling with ARFIMA processes

In order to fit a ARFIMA model to the data $\{\tilde{X}_n : n = 1, 2, \ldots, 9939\}$, first, we study the memory structure of the series.
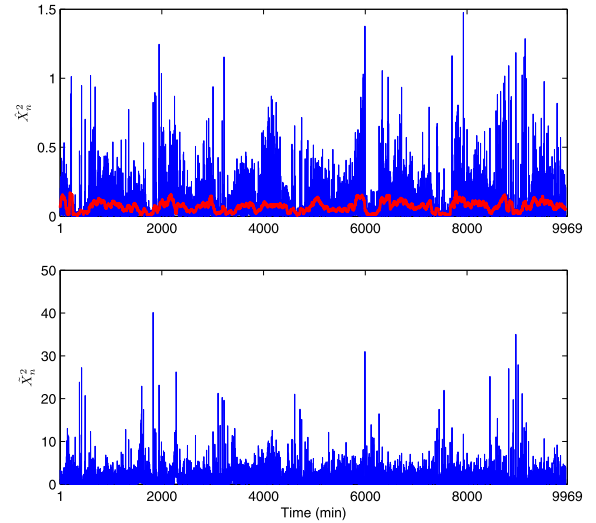

Fig. 6. (Top panel) Squared UMTS data $\{\hat{X}_n^2 : n = 31, 32, \ldots, 9969\}$ with removed seasonality, and sample variances $\{S_n^2 : n = 1, 2, \ldots, 9939\}$ (red line). (Bottom panel) Squared UMTS data $\{\tilde{X}_n^2 = \hat{X}_{n+30}^2/S_n^2 : n = 1, 2, \ldots, 9939\}$ with removed seasonality and smoothed variance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
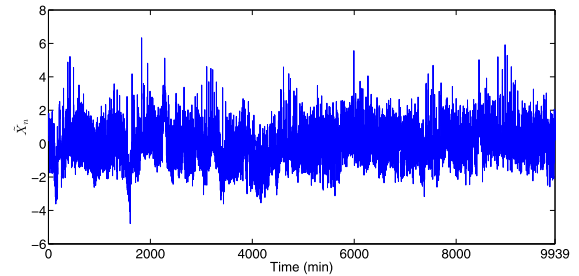

Fig. 7. UMTS data $\{\tilde{X}_n = \hat{X}_{n+30}/\sqrt{S_n^2} : n = 1, 2, \ldots, 9939\}$ with removed seasonality and smoothed variance.
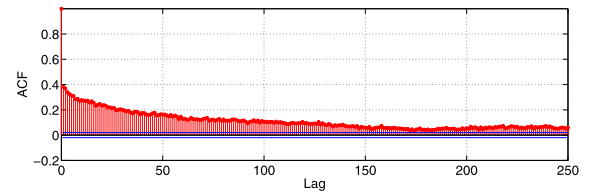

Fig. 8. Autocorrelation function of the series $\{\tilde{X}_n : n = 1, 2, \ldots, 9939\}$.

To this end we apply three methods described in Section 3.1 of estimating the memory parameter d, namely the RS, MRS and sample MSD. We obtained the following estimates: $\hat{d}_{RS} = 0.33$, $\hat{d}_{MRS} = 0.33$ and $\hat{d}_{MSD} = 0.31$. All estimators returned similar positive values, which indicates the long memory property. This can be also observed from the autocorrelation function, see Fig. 8. Next, we fractionally differentiate the series $\{\tilde{X}_n : n = 1, 2, \ldots, 9939\}$ according to the fractional differencing algorithm presented in Section 3.2. We apply such procedure with memory parameter $d = 0.32$, which is the average value of obtained estimators $\hat{d}_{RS}$ $\hat{d}_{MRS}$ and $\hat{d}_{MSD}$. After fractional differencing transformation we get the series $\{Y_n : n = 1, 2, \ldots, 4970\}$ of length 4970, see Fig. 9. To illustrate that the fractional differencing removed the dependence between far distant observations we present the ACF and PACF of the new series $\{Y_n : n = 1, 2, \ldots, 4970\}$, see Fig. 10. Moreover, we calculate the estimators of the memory parameter d, for these data, namely $\hat{d}_{RS} = 0.06$, $\hat{d}_{MRS} = -0.01$ and $\hat{d}_{MSD} = 0.01$. All values are close to zero which indicates the lack of dependence.
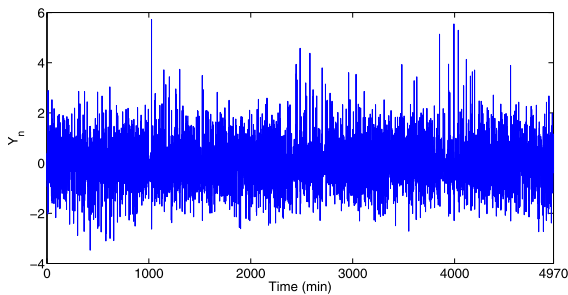
Fig. 9. Series $\{Y_n : n = 1, 2, \ldots, 4970\}$ obtained by fractional differencing with $d = 0.3232$.
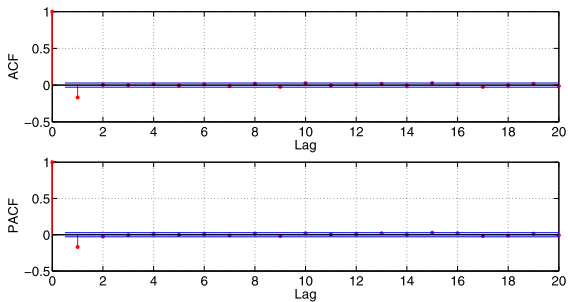

Fig. 10. (Top panel) Autocorrelation function and (bottom panel) partial autocorrelation function of the series $\{Y_n : n = 1, 2, \ldots, 4970\}$. There is no significant dependence between observations for lags greater than 1.

Table 2
The AICC statistics for different model orders (p, q).

Model order (p, q)	Value of $AICC(p, q)$
$(p, q) = (0, 0)$	$AICC(0, 0) = 14160.7$
$(p, q) = (1, 0)$	$AICC(1, 0) = 14020.2$
$(p, q) = (0, 1)$	$AICC(0, 1) = 14017.1$
$(p, q) = (1, 1)$	$AICC(1, 1) = 14019$

Table 3
The fitted ARFIMA (to the original data) and ARMA (to the differenced data) models.

Model order	β_n	ARFIMA	ARMA
$(p, q) = (1, 0)$	(d, ψ_1)	$(0.34, -0.17)$	$(-0.01, -0.16)$
$(p, q) = (0, 1)$	(d, θ_1)	$(0.39, 0.23)$	$(0.01, 0.19)$
$(p, q) = (1, 1)$	(d, ψ_1, θ_1)	$(0.44, 0.17, 0.44)$	$(0.02, 0.03, 0.22)$

We can treat the series $\{Y_n : n = 1, 2, \ldots, 4970\}$ as an ARMA series. To determine the model order (p, q) we compute the bias-corrected version of the Akaike criterion, defined in (15). The results for different model order are presented in Table 2. The applied Akaike criterion favors the model order $(0, 1)$, so the series $\{Y_n : n = 1, 2, \ldots, 4970\}$ can be described by the ARMA(0, 1) model (which is simply MA(1)). Equivalently, the series $\{\tilde{X}_n : n = 1, 2, \ldots, 9939\}$ can be treated as a ARFIMA(0, d, 1) time series.

Because both autocorrelation and partial autocorrelation functions indicate some relevant dependence for lag 1 (see Fig. 10), we decide to consider not only the model order $(p, q) = (0, 1)$, but also $(p, q) \in \{(1, 0), (1, 1)\}$. For considered model orders, we can apply the ARFIMA estimator β_n defined in (18). Results of the estimation procedure for different model orders are presented in Table 3.

Moreover, we apply the ARFIMA estimator β_n to the fractionally differenced ARMA series $\{Y_n : n = 1, 2, \ldots, 4970\}$. We obtain that d is close to zero in each case (see Table 3), which implies that there is no long memory in the data $\{Y_n : n = 1, 2, \ldots, 4970\}$
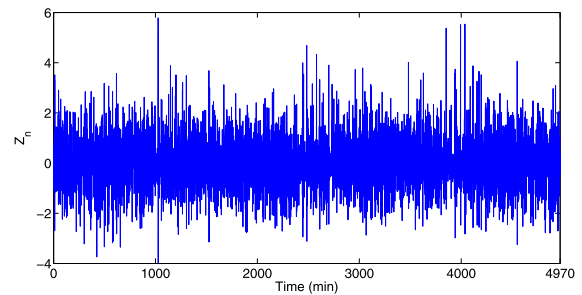

Fig. 11. Residual series $\{Z_n : n = 1, 2, \ldots, 4970\}$ from the fitted ARMA(1, 1) model with $\psi_1 = 0.17$ and $\theta_1 = 0.44$.

which justifies ARMA as an appropriate model. Simultaneously we get coefficients ψ_1 and θ_1 which are quite close to the values obtained for the ARFIMA series $\{\tilde{X}_n : n = 1, 2, \ldots, 9939\}$. Therefore, we can conclude that the proposed fractional differencing algorithm works well establishing a correct link between ARFIMA and simpler ARMA models. Summarizing, we suspect that an appropriate model for the series $\{\tilde{X}_n : n = 1, 2, \ldots, 9939\}$ is one of proposed ARFIMA processes. We analyze the three candidates in next sections.

5.3. Residual analysis

The correctness of the fitted model should be verified by checking randomness of the residuals. For the data $\{\tilde{X}_n : n = 1, 2, \ldots, 9939\}$, we consider here ARFIMA models with order up to $(p, q) = (1, 1)$. Hence, the data after fractional differencing, so the series $\{Y_n : n = 1, 2, \ldots, 4970\}$, can be treated as an ARMA process satisfying the equations

$$Y_n - \psi_1 Y_{n-1} = Z_n + \theta_1 Z_{n-1}, \tag{19}$$

or equivalently

$$(1 - \psi_1 B)Y_n = (1 + \theta_1 B)Z_n$$

for $n = 1, 2, \ldots, 4970$, where B is the backward operator defined in Section 1. The formula (19) defines the ARMA(1, 1) process. The ARMA(1, 0) and ARMA(0, 1) correspond to the cases when $\theta_1 = 0$ and $\psi_1 = 0$ respectively. Because the proposed ARMA processes are invertible (see the coefficients in Table 3), Eq. (19) can be rewritten as

$$\left[\frac{1 - \psi_1 B}{1 + \theta_1 B}\right]Y_n = [1 - \psi_1 B]\left[\frac{1}{1 + \theta_1 B}\right]$$

$$= [1 - \psi_1 B]\left[\sum_{j=0}^{\infty} (-\theta_1 B)^j\right]Y_n = Z_n.$$

It is obvious that the infinite series in the above formula has to be replaced by a finite one, because we have the finite series $\{Y_n : n = 1, 2, \ldots, 4970\}$. The innovations series $\{Z_n : n = 1, 2, \ldots, 4970\}$ of the ARMA(1, 1) case are presented in Fig. 11.

The first step to check independence of residuals for the proposed three models are plots of autocorrelation and partial autocorrelation functions. If data are independent, this property is also preserved for transformed observations like squared or absolute values. Therefore, in Fig. 12 we present autocorrelation and partial autocorrelation functions for original innovations as well as for their transformations. From Fig. 12 we can imply that only the residuals from the model of order $(p, q) = (1, 0)$ seem to be independent. This hypothesis will be verified by applying statistical tests described in Section 4.1 for the residuals of the ARFIMA(1, d, 0) model.
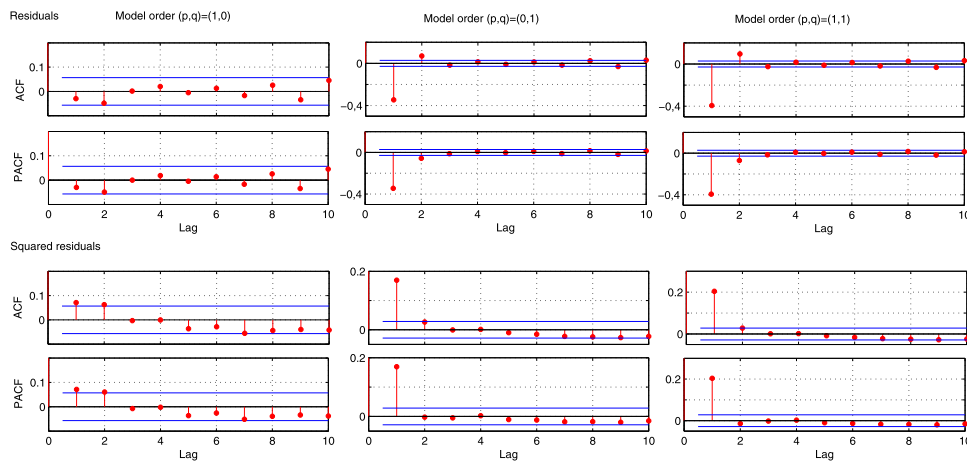
Fig. 12. Autocorrelation and partial autocorrelation functions of residuals (top raw) and squared residuals (bottom raw). The first column corresponds to the model of order $(p, q) = (1, 0)$, the second to $(p, q) = (0, 1)$ and the third to $(p, q) = (1, 1)$.
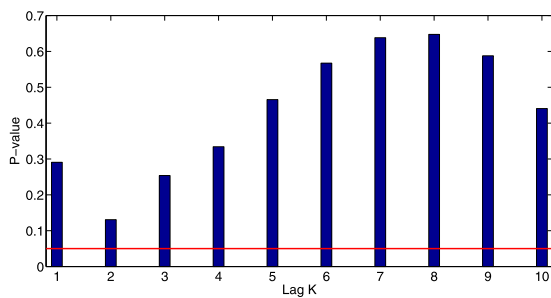

Fig. 13. The *p*-values of Ljung-Box test for different lags $K = 1, 2, \ldots, 10$. The red line corresponds the confidence level $\alpha = 0.05$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We start testing of randomness with the Ljung-Box test. We apply the test for different lags $K = 1, 2, \ldots, 10$ to check the strength of dependence in the group of observations distant by K. We obtain ten *p*-values, each significantly higher than the confidence level $\alpha = 0.05$, see Fig. 13. Hence, statistically we cannot reject the i.i.d. hypothesis for the innovations of the fitted ARFIMA(1, *d*, 0) model. We draw the same conclusion from the turning points test since the obtained *p*-value is 0.09. Therefore, we propose the following ARFIMA for the UMTS data:

$$(1 - B)^{0.34}(X_t + 0.17X_{t-1}) = Z_t. \tag{20}$$

The next task is to discover the underlying distribution for innovations of the ARFIMA(1, *d*, 0). To this end we performed several

statistical test described in Section 4.2. We tried to fit four probability distribution introduced in Section 4.2, namely Gaussian, α-stable, normal-inverse Gaussian (NIG) and tempered α-stable. In all cases statistical tests produced very low *p*-values, below the 0.05 significance level. Hence, none of proposed distribution fits the analyzed residuals. Therefore, we perform further analyzes on empirical distribution function of the residuals.

5.4. Backtesting

In this section we assess the accuracy of fitted models predictors using existing historical data. We start by selecting a threshold date which divides the data into two parts, where the latter consists of 500 observations. We calculate now one step forecasts based on the first part of the data, for the proposed ARFIMA(1, *d*, 0) model introduced in Section 4.3, and compare it with observations $\{\tilde{X}_n\}$. The result are presented in the left panel of Fig. 14 with corresponding 90% confidence prediction intervals. We can see the fitted model performs quite well. Finally, we revert the preliminary transformation on the data, namely removal of seasonality and fluctuating volatility and present the comparison for the original data, see the right panel of Fig. 14.

6. Conclusions

The ARFIMA process can serve as a universal and simple discrete time model for fractional dynamics of empirical data and the celebrated FBM and FSM form the limiting case of ARFIMA [7]. It offers a lot of flexibility in modeling of long (power-like) and
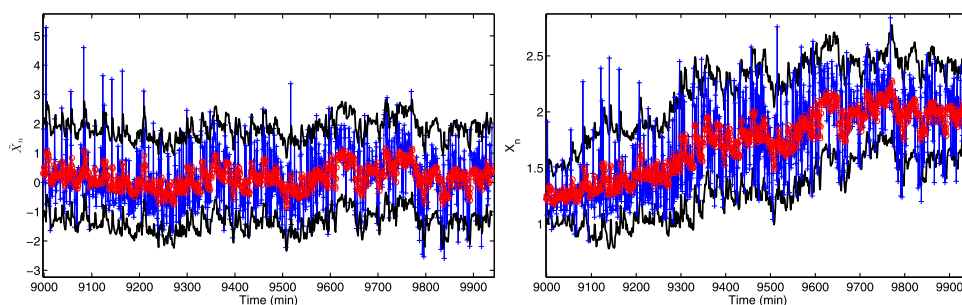

Fig. 14. (Left panel) One step prediction of the proposed FIMA(1, *d*, 0) model (red circles) and the series $\tilde{X}_n$ (blue crosses) for the last 500 observations with 90% confidence interval (black lines). (Right panel) The one step prediction of the proposed ARFIMA(1, *d*, 0) model (red circles) with added seasonality and volatility components and historical data X_n (blue crosses) for the 500 observations with 90% confidence interval (black lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

short (exponential) dependencies by choosing the memory parameter d and appropriate autoregressive and moving average coefficients. Modeling with ARFIMA processes also allows for taking into account different light and heavy-tailed distributions. Hence, they are well tailored to different empirical data.

In this paper we presented a scheme for identification and validation of ARFIMA time series with noise belonging to the domain of attraction of stable law. The identification part incorporates a new fractional differencing algorithms which enables to arrive at ARMA times times which are easier to handle and consequently accurately estimate the order of the model. To apply the fractional differencing a preliminary estimate of the parameter d is needed. To this end we used the low variance mean-squared displacement estimator [7]. The ARFIMA parameter estimation applies a version of Whittle's estimator which was proposed in [61] for both negative and positive d's. The validation part involves standard diagnostics of the residuals based on statistical tests for independence, distribution fitting to residuals and backtesting procedure.

The usefulness of the identification and validation scheme was illustrated on an UMTS data set. The electromagnetic field intensity was measured in Wroclaw in an urban area in 2011. The data clearly possessed a seasonal trend with a daily period which was related to a daily pattern of the telephone calls. After removing the daily seasonality the autocorrelation function still displayed a non-stationary behavior. Hence, we studied the volatility of the data. A seasonal fluctuating variance was observed in a squared data plot which was removed by subtracting the sample moving variance. Finally, we showed that a long memory ARFIMA(1, 0.34, 0) model describes the transformed data well. This was justified by means of visual and statistical tests on residuals. The distribution of residuals could not be identified. Such ARFIMA process can be simulated by generating the noise from its empirical distribution function. We believe that the proposed ARFIMA methodology can be also applied to other telecommunication data. This would allow to better predict hazardous levels of the electromagnetic field intensity.

Finally, we note that the identification scheme can be also applied to more general ARFIMA processes with generalized autoregressive conditional heteroskedasticity (GARCH) noise. Such processes are called ARFIMA-GARCH [84]. This extension can be very useful for modeling so-called transient diffusion phenomena, where the diffusion coefficient fluctuates randomly in time [85–87]. For the UMTS data the variance changes in time but the changes have a seasonal pattern hence the ARFIMA-GARCH model does not provide a better fit.

Acknowledgment

References

[1] Klafter J, Lim S, Metzler R, editors. Fractional dynamics: recent advances. Singapore: World Scientific; 2012.
[2] Meerschaert MM, Sikorskii A. Stochastic models for fractional calculus. De gruyter studies in mathematics, vol. 43; 2012.
[3] Metzler R, Klafter J. The random walks guide to anomalous diffusion: a fractional dynamics approach. Phys Rep 2000;339:1–77.
[4] Beran J. Statistics for long-memory processes. New York: Chapman & Hall; 1994.
[5] Doukham P, Oppenheim G, Taqqu MS. Theory and applications of long range dependence. Boston: Birkhäuser; 2003.
[6] Samorodnitsky G. Long range dependence. Found Trends Stoch Syst 2006;1:163–257.
[7] Burnecki K, Weron A. Algorithms for testing of fractional dynamics: a practical guide to ARFIMA modelling. J Stat Mech 2014:P10036.
[8] Cajueiro DO, Tabak BM. Time-varying long-range dependence in US interest rates. Chaos Solitons Fractals 2007;34:360–7.
[9] Souzo SR, Tabak BM, Cajueiro DO. Long memory testing for fed funds futures contracts. Chaos Solitons Fractals 2008;37:180–6.
[10] Cajueiro DO, Tabak BM. Testing for long-range dependence in the brazilian term structure of interest rates. Chaos Solitons Fractals 2009;40:1559–73.
[11] Barkoulas JT, Barilla AG, Wells W. Long-memory exchange rate dynamics in the euro era. Chaos Solitons Fractals 2016;86:92–100.
[12] Lo AW. Long-term memory in stock market prices. Econometrica 1991;59:1279–313.
[13] Lo AW. Fat tails, long memory, and the stock market since the 1960s. Econ Notes 2001;26:219–52.
[14] Burnecki K. Self-similar processes as weak limits of a risk reserve process. Probab Math Statist 2000;20:261–72.
[15] Baillie RT. Long memory processes and fractional integration in econometrics. J Econ 1996;73:5–59.
[16] Beran J, Sherman R, Taqqu MS, Willinger W. Long-range dependence in variable-bit-rate video traffic. IEEE Trans Commun 1995;43:1566–79.
[17] Norros I. On the use of fractional brownian motion in the theory of connectionless networks. IEEE J Sel Areas Commun 1995;13:953–62.
[18] Willinger W, Taqqu MS, Sherman R, Wilson D. Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. IEEE/ACM Trans Net 1997;5:71–96.
[19] Karagiannis T, Molle M, Faloutsos M. Long-range dependence ten years of internet traffic modeling. IEEE Internet Comput 2004;8:57–64.
[20] Coulon M, Chabert M, Swami A. Detection of multiple changes in fractional integrated ARMA processes. IEEE Trans Signal Process 2009;57:48–61.
[21] Stoev SA, Michailidis G, Taqqu MS. Estimating heavy-tail exponents through max self "similarity". IEEE Trans Inf Theory 2011;57. 1651–1636
[22] Painter S.. Long-range dependence in the subsurface: Empirical evidence and simulation methods. Invited paper at the American Geophysical Union 1998 Fall Meeting.
[23] Varotsos C, Kirk-Davidoff D. Long-memory processes in ozone and temperature variations at the region 60° S-60° N. Atmos Chem Phys 2006;6:4096–100.
[24] Alvarez-Lacalle E, Dorow B, Eckmann JP, Moses E. Hierarchical structures induce long-range dynamical correlations in written texts. PNAS 2006;103:7956–61.
[25] Karmeshu D, Krishnamachari A, et al. Sequence variability and long-range dependence in DNA: an information theoretic perspective. In: Pal N, et al., editors. Neural information processing. Springer: Berlin; 2004. p. 1354–61.
[26] Peng CK, Mietus J, Hausdorff J, Havlin S, Stanley HE, Goldberger A. Long-range anticorrelations and non-Gaussian behaviour of the heartbeat. Phys Rev Lett 1993;70:1343–6.
[27] Szymanski J, Weiss M. Elucidating the origin of anomalous diffusion in crowded fluids. Phys Rev Lett 2009;103:038102.
[28] Burnecki K, Weron A. Fractional lévy stable motion can model subdiffusive dynamics. Phys Rev E 2010;82:021130.
[29] Kepten E, Bronshtein I, Garini Y. Ergodicity convergence test suggests telomere motion obeys fractional dynamics. Phys Rev E 2011;83:041919.
[30] Burnecki K. FARIMA processes with application to biophysical data. J Stat Mech 2012:P05015.
[31] Burnecki K, Muszkieta M, Sikora G, Weron A. Statistical modelling of subdiffusive dynamics in the cytoplasm of living cells: a FARIMA approach. EPL 2012;98:10004.
[32] Burnecki K, Kepten E, Garini Y, Sikora G, Weron A. Estimating the anomalous diffusion exponent for single particle tracking data with measurement errors - an alternative approach. Sci Rep 2015;5:11306.
[33] Stanislavsky A, Burnecki K, Magdziarz M, Weron A, Weron K. FARIMA modeling of solar flare activity from empirical time series of soft X-ray solar emission. Astrophys J 2009;693:1877–82.
[34] Stoev S, Taqqu MS. Simulation methods for linear fractional stable motion and FARIMA using the fast fourier transform. Fractals 2004;12:95–121.
[35] Guo H, Lim CY, Meerschaert MM. Local whittle estimator for anisotropic random fields. J Multivariate Anal 2009;100:993–1028.
[36] Graves T, Gramacy RB, Franzke CL, Watkins NW. Efficient bayesian inference for natural time series using ARFIMA processes. Nonlin Processes Geophys 2015;22:679–700.
[37] Lanoiselée Y, Grebenkov DS. Revealing nonergodic dynamics in living cells from a single particle trajectory. Phys Rev E 2016;93:052146.
[38] Burnecki K, Sikora G, Weron A. Fractional process as a unified model for subdiffusive dynamics in experimental data. Phys Rev E 2012;86:041912.
[39] Meerschaert MM, Scheffler HP. Limit theorems for continuous time random walks with infinite mean waiting times. J Appl Probab 2004;41:623–38.
[40] Magdziarz M, Weron A. Ergodic properties of anomalous diffusion processes. Ann Phys 2011;326:2431–43.
[41] Janczura J, Weron A. Ergodicity testing for anomalous diffusion: small sample statistics. J Chem Phys 2015;142:144103.
[42] Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data. SIAM Rev 2009;51:661–703.
[43] Nolan J. Stable distributions. Models for heavy-Tailed data. Boston: Birkhäuser; 2007.
[44] Rachev S, Mittnik S. Stable paretian models in finance. New York: Wiley and Sons; 2000.
[45] Tankov P, Cont R. Financial modelling with jump processes. London: Chapman and Hall/CRC Financial Mathematics Series; 2003.

[46] Rachev S, editor. Handbook of heavy tailed distributions in finance. Amsterdam: Elsevier Science B.V.; 2003.

[47] Weron A, Weron R. Fractal market hypothesis and two power-laws. Chaos Solitons Fractals 2000;11:289–96.

[48] Li H, Wu M, Wang XT. Fractional-moment capital asset pricing model. Chaos Solitons Fractals 2009;42:412–21.

[49] Burnecki K, Wyłomańska A, Chechkin A. Discriminating between light- and heavy-tailed distributions with limit theorem. PLoS ONE 2015;10(12):e0145604.

[50] Gonchar VY, Chechkin AV, Sorokovoi EL, Chechkin VV, Grigoreva LI, Volkov ED. Stable lévy distributions of the density and potential fluctuations in the edge plasma of the u-3m torsatron. Plasma Phys Rep 2003;29:380–90.

[51] Burnecki K, Wyłomańska A, Beletskii A, Gonchar V, Chechkin A. Recognition of stable distribution with lévy index alpha close to 2. Phys Rev E 2012;85:056711.

[52] Joelson M, Golder J, Beltrame P, Néel MC, Pietro LD. On fractal nature of groundwater level fluctuations due to rainfall process. Chaos Solitons Fractals 2016;82:103–15.

[53] Ditlevsen PD. Observation of alpha-stable noise induces millenial climate changes from an ice record. Geophys Res Lett 1999;26:1441–4.

[54] Ditlevsen PD. Anomalous jumping in a double-well potential. Phys Rev E 1990;60:172–9.

[55] Peng CK, Mietus J, Hausdorff JM, Havlin S, Stanley HE, Goldberger AL. Long-range anticorrelations and non-gaussian behavior of the heartbeat. Phys Rev Lett 1993;70:1343–6.

[56] Nikias CL, Shao M. Signal processing with alpha-stable distributions and applications. New York: Wiley; 1995.

[57] Mantegna RN, Stanley HE. An introduction to econophysics. Cambridge: Cambridge University Press; 2000.

[58] Burnecki K, Gajda J, Sikora G. Stability and lack of memory of the returns of the hang seng index. Phys A 2011;390:3136–46.

[59] Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. Nature 2006;439:462–5.

[60] Sims DW, Southall EJ, Humphries NE, Hays GC, Bradshaw CJA, Pitchford JW. Scaling laws of marine predator search behaviour. Nature 2008;451:1098–103.

[61] Burnecki K, Sikora G. Estimation of FARIMA parameters in the case of negative memory and stable noise. IEEE Trans Signal Process 2013;61:2825–35.

[62] Granger CWJ, Joyeux R. An introduction to long–memory time series and fractional differencing. J Time Series Anal 1980;1(1):15–30.

[63] Hosking JRM. Fractional differencing. Biometrika 1981;68(1):165–76.

[64] Samorodnitsky G, Taqqu MS. Stable non–aussian random processes: stochastic models with infinite variance,. London: Chapman & Hall; 1994.

[65] Kokoszka PS, Taqqu MS. Fractional ARIMA with stable innovations. Stochastic Process Appl 1995;60(1):19–47.

[66] Geweke J, Porter-Hudak S. The estimation and application of long memory time series models. J Time Ser-Anal 1983;4:221–38. 1983

[67] Hurst HE. Long-term storage capacity of reservoirs. Trans Am Soc Civ Eng 1951;116:770–99.

[68] Taqqu MS, Teverovsky V. On estimating the intensity of long–range dependence in finite and infinite variance time series. A practical guide to heavy tails: statistical techniques and applications. Adler R, Feldman R, Taqqu MS, editors. Boston: Birkhäuser; 1998.

[69] Taqqu MS, Teverovsky V, Willinger W. A critical look at lo's modified RS statistic. J Statist Plann Inference 1999;80:211–27.

[70] Frigo M, Johnson SG. A modified split-radix FFT with fewer arithmetic operations. IEEE Trans Signal Process 2007;55(1):111–19.

[71] Liu J, Shu Y, Zhang L, Xue F. Traffic modeling based on FARIMA models. In: Electrical and computer engineering, 1999 IEEE canadian conference, vol. 1; 1999. p. 162–7.

[72] McCarthy J, DiSario R, Saraoglu H. A recursive algorithm for fractionally differencing long data series. J Mod App Stat Meth 2003;2:272–8.

[73] Xiu J, Jin Y. Empirical study of ARFIMA model based on fractional differencing. Phys A 2007;377:138–52.

[74] Hannan EJ. The asymptotic theory of linear time-series models. J Appl Probab 1973;10:130–45. Corrections, ibid. vol. 10 (1973), p. 913

[75] Brockwell P, Davis R. Introduction to time series and forecasting. New York: Springer-Verlag; 2002.

[76] Kokoszka PS, Taqqu MS. Parameter estimation for infinite variance fractional ARIMA. Ann Statist 1996;24(5):1880–913.

[77] Ljung GM, Box GEP. On a measure of lack of fit in time series models. Biometrika 1978;65:297–303.

[78] Ross S. Simulation. San Diego: Academic Press; 2002.

[79] Barndorff-Nielsen OE. Normal inverse gaussian processes and the modelling of stock returns. Research Report 300. Department of Theoretical Statistics, University of Aarhus; 1995.

[80] Weron R.. Computationally intensive value at risk calculations. In: Gentle J.E., Haerdle W., Mori Y., editors. Handbook of computational statistics: concepts and methods. Berlin: Springer; p. 911.

[81] Rosiski J. Tempering stable processes. Stochastic Process Appl 2007;117(6):677–707.

[82] Baeumer B, Meerschaert MM. Tempered stable lvy motion and transient super-diffusion. J Comput Appl Math 2010;233:2438–48.

[83] Kokoszka PS. Prediction of infinite variance fractional ARIMA. Probab Math Statist 1996;16:65–83.

[84] Ling S, Li WK. On fractionally integrated autoregressive moving-average time series models with conditional heteroscedasticity. J Amer Statist Assoc 1997;92:1184–94.

[85] Bronstein I, et al. Transient anomalous diffusion of telomeres in the nucleus of mammalian cells. Phys Rev Lett 2009;103:018102.

[86] Godec A, Bauer M, Metzler R. Collective dynamics effect transient subdiffusion of inert tracers in flexible gel networks. New J Phys 2014;16:092002.

[87] Manzo C, et al. Weak ergodicity breaking of receptor motion in living cells stemming from random diffusivity. Phys Rev X 2015;5:011021.