Modified cumulative distribution function in application to waiting time analysis in the continuous time random walk scenario

Rafał Połoczański¹, Agnieszka Wyłomańska¹, Monika Maciejewska², Andrzej Szczurek² and Janusz Gajda¹

¹ Faculty of Pure and Applied Mathematics, Hugo Steinhaus Center, Wroclaw University of Science and Technology, Poland
² Institute of Air Conditioning and District Heating, Wroclaw University of Science and Technology, Poland

E-mail: rafal.poloczanski@pwr.edu.pl, agnieszka.wylomanska@pwr.edu.pl, monika.maciejewska@pwr.edu.pl, andrzej.szczurek@pwr.edu.pl and janusz.gajda@pwr.edu.pl

Received 20 May 2016, revised 19 October 2016
Accepted for publication 27 October 2016
Published 9 December 2016

Abstract
The continuous time random walk model plays an important role in modelling of the so-called anomalous diffusion behaviour. One of the specific properties of such model is the appearance of constant time periods in the trajectory. In the continuous time random walk approach they are realizations of the sequence called waiting times. In this work we focus on the analysis of waiting time distribution by introducing novel methods of parameter estimation and statistical investigation of such a distribution. These methods are based on the modified cumulative distribution function. In this paper we consider three special cases of waiting time distributions, namely \(\alpha\)-stable, tempered stable and gamma. However, the proposed methodology can be applied to broad set of distributions—in general it may serve as a method of fitting any distribution function if the observations are rounded. The new statistical techniques are applied to the simulated data as well as to the real data of \(\text{CO}_2\) concentration in indoor air.

Keywords: waiting time, continuous time random walk, distribution, estimation

(Some figures may appear in colour only in the online journal)
1. Introduction

It is widely recognized that many real empirical data exhibit very specific behaviour, namely one can observe time periods where the process stays at the same level. This special property may indicate that the theoretical model behind the data is based on the so-called continuous time random walk (CTRW) [1]. This model, in the classical version, is constructed as a sum of independent random variables (called later jumps) from the same distribution. The number of jumps is a realization of the process based on the so-called waiting time sequence, therefore the distribution of the CTRW model is defined through the jumps as well as waiting time distributions.

In the CTRW scheme, originally introduced in [2], the waiting times between the jumps are not constant, as in the standard random walk, but they are random variables governed by a probability law. As a consequence, the CTRW model is a natural description of transport in crowded environments and complex systems. It is worth mentioning that the limit distribution of the CTRW scheme can be also formulated within the framework of the fractional Fokker–Planck equation [3]. The limit of the properly scaled CTRW is a process (called subordinated process) which is a superposition of two systems: one described by Langevin-type equations and another called the inverse subordinator [4]. Over the last years subordinated processes have appeared in the literature more and more often [5–7].

The CTRW model is one of the most important members of the system of anomalous diffusion processes [8–10]. Anomalous diffusion models have found many practical applications. They have been used in a variety of physical systems, including charge carrier transport in amorphous semiconductors [11–13], transport in micelles [14], intracellular transport [15], motion of mRNA molecules inside Escherichia coli cells [16] or the motion of telomeres in the nucleus of mammalian cells [17]. Other interesting applications to physical data modelling can be found in [18–20]. In addition, the behaviour corresponding to CTRW scenario can be observed in stock prices or interest rate data [21, 22] as well as in the time series of indoor air quality parameters [23–25].

One of the most important issues that arise in the analysis of the CTRW model is the description of waiting times. A knowledge of waiting time distribution allows one to conclude on the properties of the whole process. The most popular distribution of constant time periods is the $\alpha$-stable one [5, 22]. However, recent developments indicate that another non-negative infinitely divisible distribution can also be used to model the observed waiting times [26–29].

There have been several successful attempts to apply waiting time analysis in many areas of research and practice. In this work we demonstrate its usefulness in the domain of measurements. A continuous monitoring of various quantities is used in many applications in preference to other types of measurements [30, 31]. Usually, it provides a great number of discrete data. They have to be transmitted, verified and analysed [32]. These operations consume a significant amount of time and energy, therefore they pose a serious problem in many monitoring systems, especially in wireless networks [33]. Additionally, continuous monitoring may be a source of useless data. This is a result of the measurement characteristics of sensors (e.g. inappropriate accuracy and sensitivity). These disadvantages can be reduced by the application of a suitable sampling procedure [34]. Sampling indicates how many data need to be collected and how often it should be done in order to gain access to the required information. In this work, it is demonstrated that the statistical analysis of waiting times for an essential change of CO$_2$ concentration inside room may be applied to specify the most appropriate time interval between consecutive measurements. In this way, the sampling rate (frequency of sampling) is determined. It was assumed that the sampling procedure could reduce the number of data, but the information content cannot be significantly reduced.
The main goal of the paper is to introduce a new method of parameter estimation and statistical investigation of waiting time distribution in the CTRW scenario. The most commonly used method to identify the correct distribution of constant time periods is visualization of the empirical cumulative distribution function and its comparison with a fitted parametric distribution [35]. There are also methods for comparison between empirical and theoretical characteristic functions [36]. However, testing procedures which would allow one to identify the correct class of distribution of waiting time observations are rarely considered in the literature. The difficulty arises from the discretization of the time intervals, which is a natural consequence of time-discretization of measuring devices. Moreover, as we show in section 3, standard estimation procedures might lead to under- or overestimation of parameters. The proposed methods of parameter estimation and statistical testing of proper distributions are based on the modified cumulative distribution function. Here we propose a novel technique that can be applied to a broad set of distributions—in general it may serve as a method of fitting of any distribution function if the observations are rounded. In our analysis we consider three waiting time distributions, namely α-stable, tempered stable and gamma, as the most commonly used to describe constant time periods.

The rest of the paper is organized as follows: in section 2 we introduce the CTRW model with three possible waiting time distributions, namely α-stable, tempered stable and gamma, and indicate the main properties of the considered models. Next, in section 3 we introduce the modified cumulative distribution function which serves as a basis for parameter estimation and testing of the proper distribution of constant time periods. In section 4 we present how the introduced methods work for simulated data from CTRW, while in section 5 we analyse the real time series of CO₂ concentration in indoor air in the context of the presented methodology. The final section concludes the paper.

2. Model

In this section we present the CTRW methodology [2] that was first introduced to describe the motion of a particle which arrives at a position, waits for a random time and then jumps randomly to the new position. Thus such a motion is determined by waiting times and the distribution of jumps. CTRW models are nowadays well established and popular mathematical objects, particularly attractive in the description of the so-called anomalous diffusion phenomenon [3, 37]. It is an interesting observation that many CTRWs converge under suitable assumptions to the so-called subordinated processes, i.e. processes where physical time is replaced by another stochastic process [38]. For instance a CTRW model with a power-law heavy-tailed α-stable waiting times and finite second-moment jump distribution converges to the subordinated Brownian motion [39] \( B(S^\alpha_{\alpha}(t)) \). Here \( B(\cdot) \) is a Brownian motion and \( S^\alpha_{\alpha}(\cdot) \) is the so-called inverse α-stable subordinator defined in (7). However, one can name many more examples of subordinated processes and their different applications, both with regard to outer and inner processes. To name only few, we mention subordinated fractional Brownian motion [40, 41], the subordinated α-stable Lévy process, where the authors considered a tempered stable subordinator [42], and inverse gamma subordinated Brownian motion [29].

In this paper we analyse a real data set of discrete observations for both time and scale; hence it is reasonable to consider it as a proper model system based on the CTRW approach. In such a classical CTRW setting one considers the following stochastic process:

\[
Y(t) = \sum_{i=1}^{N(t)} X_i, \tag{1}
\]
where $N(t)$ is a counting process defined as

$$
N(t) = \max \left\{ k \geq 0 : \sum_{i=1}^{k} T_i \leq t \right\}.
$$

Here $T_i, i = 1, 2, \ldots$ form a sequence of positive independent, identically distributed (IID) random variables which can be seen as waiting times between consecutive jumps $X_i$. We assume that the sequences $\{T_i\}_{i=1}^{\infty}$ and $\{X_i\}_{i=1}^{\infty}$ are independent. In figure 1 we present an exemplary trajectory of the CTRW process.

In this paper we consider three particular distributions of waiting times for the process defined in (1), namely $\alpha$-stable, tempered stable and gamma. We denote such random variables as $T_i^S$, $T_i^T$ and $T_i^G$, respectively. In the case when the sequence $\{T_i^S\}$ constitutes a sample of IID random variables with $\alpha$-stable distribution, for each $i$ the random variable $T_i^S$ has the following Laplace transform [43]:

$$
\mathbb{E}(e^{-z T_i^S}) = e^{-\sigma \alpha \alpha},
$$

where $\sigma > 0$ is a scale parameter and $\alpha \in (0, 1)$. One can find the properties of $\alpha$-stable distributed random variables in [44], for example. We only mention that the $\alpha$-stable family has two important properties. First, a sum of two independent $\alpha$-stable random variables with the same $\alpha$ parameter is also $\alpha$-stable distributed. Second, tails of the stable distribution are governed by power-law behaviour.

In the case of tempered stable waiting times, the $T_i^T$ random variable has the following Laplace transform [26, 45]:

$$
\mathbb{E}(e^{-z T_i^T}) = e^{\sigma (\lambda^\alpha - (\lambda + \lambda^\alpha))},
$$

for some parameters $\sigma, \lambda > 0$ and $\alpha \in (0, 1)$. One observes that by taking $\lambda = 0$ tempered law reduces to $\alpha$-stable one. Tempered stable distributions are particularly important in applications due to the fact that they can be at the same time close to $\alpha$-stable distributions and possess finite moments of all orders. Thus tempered stable processes become very popular in various applications. The probability density function (PDF) of a tempered stable distributed
random variable has no explicit form; however, from the above Laplace transform one can infer relation between PDFs of pure $\alpha$-stable ($g_{TS}$) and tempered stable ($g_{TT}$) distributions, namely [24]

$$g_{TT}(x) = e^{-\lambda x} x^\alpha g_{TS}(x).$$

In the third analysed scenario of the distribution of waiting times we consider the positive gamma distribution, for which the Laplace transform has the form [29]

$$E^z_k = \frac{1}{1 + z\theta}.$$  

(5)

The gamma distributed random variable $T^G$ has the following PDF:

$$g_{TT}(x) = \frac{1}{\Gamma(k)} x^{k-1} e^{-x/\theta}, \quad x > 0,$$

(6)

where $\Gamma(k)$ is the gamma function. In the above definitions $k$ is the shape parameter while $\theta$ is the scale parameter. For $k = 1$ the gamma distributed random variable becomes exponentially distributed.

In this paper we assume that jumps possess a finite second moment, thus by applying the classical Donsker theorem [46] one can prove their convergence to Brownian motion. The same argument is valid when waiting times have finite second moment. Thus this holds in case of tempered stable and gamma waiting times in the CTRW scenario. However, when the sequence $T^S_i$ belongs to the domain of attraction of a one-sided $\alpha$-stable distribution (defined via the Laplace transform (3)) then [39]

$$n^{-1/\alpha} \sum_{i=1}^{[nt]} T^S_i \overset{d}{\to} S_\alpha(t)$$

as $n \to \infty$ for fixed $t$. Here $S_\alpha(t)$ is the $\alpha$-stable subordinator with the Laplace transform [29]

$$E^{\alpha} e^{-sT_\alpha(t)} = e^{-s^\alpha}.$$  

The notation $\overset{d}{\to}$ means ‘convergence in distribution’. In the context of the CTRW convergence let us also introduce the so-called inverse $\alpha$-stable subordinator $S^{-1}_\alpha(t)$ defined as a first passage time of $S_\alpha(t)$, i.e.

$$S^{-1}_\alpha(t) = \inf\{\tau \geq 0 : S_\alpha(\tau) > t\}.$$  

(7)

Then one can formulate the following simple fact for heavy-tailed CTRWs:

**Proposition 1.** Let $Y(t)$ be the CTRW process defined in (1). Assume that random variables $X_i$ are IID with finite mean 0 and second moment equal to 1. Moreover, the jump times $T^S_i$ belong to the domain of attraction of the one-sided $\alpha$-stable distribution (3) for some $0, \alpha \in (0, 1)$. Then

$$\frac{Y(mt)}{n^{\alpha/2}} \overset{d}{\to} B(S^{-1}_\alpha(t)),$$

(8)

where $B(\cdot)$ is a standard Brownian motion and $S^{-1}_\alpha(t)$ is the inverse $\alpha$-stable subordinator defined in (7).

Proof of this statement is a consequence of theorem 1 in [47].
3. Estimation procedure

The estimation procedure for the CTRW model parameters is divided into several steps. The specific feature of the CTRW process is the fact that it exhibits so-called ‘trapping events’ behaviour, i.e. visible constant time periods (the time intervals for which the process stays on the same level; see figure 1). Therefore in the first step in the estimation scheme we divide the analysed vector of observations into two vectors. The first one is related to the lengths of ‘trapping events’ while the second represents the vector of observations that arises after removing the ‘trapping events’. This standard procedure of analysis of CTRW models has been used in various applications (see for example [23, 48]). We estimate the parameters corresponding to the waiting time distribution on the basis of the lengths of ‘trapping events’. The whole procedure of testing and estimating the waiting time distribution is presented in the further part of this section. After estimation of the waiting time distribution we analyse the process arising as a result of removal of the ‘trapping events’, i.e. the distribution of the jumps in the CTRW scenario. The details are presented in the next sections.

At this point we would like to highlight one of the assumptions considered in this paper, namely that waiting times can be extracted from the trajectory. In real-life scenarios this is not always achievable; in [54] the authors present and analyse in details a CTRW with superimposed noise. Such behaviour is observed for instance due to thermal agitation of the environment. It is feasible to obtain a noise-free process only in certain cases and if one is able to attain a sufficiently long trajectory. However, how to subtract the noise from real data is still unknown and left for future analysis.

3.1. Analysis of waiting times

As mentioned in section 1, one of the main problems in CTRW analysis is the identification of waiting time distribution. In the literature usually it is assumed that waiting time observations follow infinitely divisible distribution, e.g. \( \alpha \)-stable, tempered stable or gamma [29]. The most commonly used method to identify the correct distribution is visualization of the empirical cumulative distribution function (CDF) on a plot with log–log scale and its comparison with a fitted parametric distribution [35]. However, testing procedures for waiting time distribution are rarely considered in the literature. The difficulty arises from the discretization of the time intervals, which is a natural consequence of time discretization of measuring devices. Furthermore, standard estimation procedures like the method of moments in the case of gamma and tempered stable distributions [40] and the McCulloch or regression method in case of the \( \alpha \)-stable distribution [49] might lead to under- or overestimation of parameters. The other methods, such as fitting the linear function in a log–log scale in the case of the \( \alpha \)-stable distribution and an exponential function for the tempered stable distribution, require the setting of arbitrary thresholds which usually depend on the data set [21].

In this paper we propose a new estimation procedure which can be applied to fit an appropriate distribution when the observations are rounded or discretized, such as waiting times until a characteristic of a process changes. In addition, we show that the procedure can be used to identify the distribution which best fits the data among other certain classes. As an example we perform a comparison between \( \alpha \)-stable, tempered stable and gamma distributions, as the ones most commonly used to describe waiting times.

The main issue is related to the fact that the exact waiting time is unknown and usually comes from a continuous distribution. For example, if we observe that a character of the
process has changed after three units of time, it is not known at which point of time the change actually happened because the correct value lies in the interval (2,4), which can be seen in figure 2. In other words, a constant time period equal to 2.5 might be classified as 2 or 3 with equal probability.

Due to these facts, we introduce a modified version of the CDF. Let $X$ be a non-negative continuous random variable with CDF $F$ and PDF $f$. We define the mass function $\tilde{f}$ with a support on natural numbers in the following way:

$$\tilde{f}(n) = \int_{n-1}^{n} f(x)(x - n + 1)dx + \int_{n}^{n+1} f(x)(n - x + 1)dx \quad \text{for } n \geq 1.$$ 

The modified CDF is defined as $\tilde{F}(x) = \sum_{i \leq x} \tilde{f}(i)$. At the points of discontinuity the function can be expressed as

$$\tilde{F}(0) = \int_{0}^{1} f(x)(1 - x)dx$$
$$\tilde{F}(n) = \tilde{F}(n - 1) + \int_{n-1}^{n} f(x)(x - n + 1)dx + \int_{n}^{n+1} f(x)(n - x + 1)dx \quad \text{for } n \geq 1,$$

where $n \in \mathbb{N}$. Since for waiting times we limit the classes of distribution functions only to those with non-negative support and assume that the observations are done in equal time intervals, it is sufficient to define $\tilde{F}$ on the set of natural numbers. However, it is straightforward to extend the definition of the function to a countable set.

Figure 2. Comparison of potential real time period versus observed value. Although the observed value is three time units, the exact value might be a short time period close to two or a long time period close to four.
The graphical interpretation of differences between the standard cumulative distribution function and the introduced modified version of it is illustrated in figure 3. It is easy to show that

\[
\tilde{F}(n) = F(n) + \int_{n}^{n+1} f(x)(n - x + 1)dx
\]

\[
= F(n) + (n + 1)(F(n + 1) - F(n)) - \int_{n}^{n+1} xf(x)dx
\]

\[
= (n + 1)F(n + 1) - nF(n) - ((n + 1)F(n + 1) - nF(n) - \int_{n}^{n+1} F(x)dx = \int_{n}^{n+1} F(x)dx.
\]

**Proposition 2.** The modified cumulative distribution function \( \tilde{F} \) can be considered as a cumulative distribution function.

**Proof.** It is easy to show that the modified cumulative distribution function \( \tilde{F} \) is bounded between 0 and 1, namely

\[
0 = \int_{n}^{n+1} 0 \leq \int_{n}^{n+1} F(x) \leq \int_{n}^{n+1} 1 = 1.
\]

Straightforward calculations show that

\[
\lim_{n \to \infty} \tilde{F}(n) = \lim_{n \to \infty} F(n) + \int_{n}^{n+1} f(x)(n - x + 1)dx \geq \lim_{n \to \infty} F(n) \ dx = 1
\]

since the upper boundary is equal to 1, we conclude that \( \lim_{n \to \infty} \tilde{F}(n) = 1 \). Since natural numbers support the mass function, it is obvious that \( \lim_{n \to \infty} \tilde{F}(n) = 0 \). Moreover, we can show the modified CDF is non-decreasing.
Finally we mention that the modified CDF $\tilde{F}$ is right-continuous, which stems from its definition.

Due to the fact that between consecutive measurements only a single change in the process can be observed, a further modification of the function $\tilde{F}$ defined in (9) is needed. We define $G$ as a rescaled modified cumulative distribution function (RMCDF) in the following way:

$$G(n) = \frac{\tilde{F}(n) - \tilde{F}(0)}{1 - \tilde{F}(0)} \quad \text{for } n \geq 0.$$  

(9)

It is obvious that $G$ still has the properties of a CDF. The comparison between both types of CDF can be seen in figure 4. The distance between the empirical CDF of waiting times can be easily observed: in the left panel the theoretical CDF of the $\alpha$-stable distribution underestimates the empirical one whereas in the right panel the theoretical CDF of the tempered stable distribution overestimates the empirical one. In both cases the rescaled modified CDF provides a reasonable fit to the data.

The proposed methodology of fitting the waiting time distribution parameters is based on the minimum distance estimation applied to a certain type of distribution. Let $K$ and $L$ denote two functions with a common support on $\mathbb{R}$; the considered distances are

- Kolmogorov–Smirnov (KS)
  $$\text{KS}(K, L) = \sup_{x \in \mathbb{R}} |K(x) - L(x)|$$

- Cramér–von Mises (CvM)
\[ \text{CvM}(K, L) = \int_{-\infty}^{\infty} (K(x) - L(x))^2 dL(x) \]

\[ \text{AD}(K, L) = \int_{-\infty}^{\infty} \frac{(K(x) - L(x))^2}{L(\alpha)(1 - L(x))} dL(x). \]

In our estimation procedure we consider the distance between the rescaled modified CDF introduced in (9) and empirical distribution function of waiting times defined by

\[ \tilde{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq t}, \]

where \( \mathbf{1}_A \) is the indicator of the set \( A \). For each class of CDFs we can find parameters which minimize the distance to the empirical distribution function, i.e. \( G_{\theta^*} \) which satisfies the following condition

\[ D(G_{\theta^*}, \tilde{F}) = \inf_{\theta \in \Theta} D(G_{\theta}, \tilde{F}), \]

where \( D \) is one of the introduced distances, KS, CvM or AD, \( \tilde{F} \) is empirical CDF introduced in (10), \( G \) is a rescaled CDF defined in (9) and \( \Theta \) is the set of parameters of a certain class of distribution functions. As already mentioned, in this paper we consider three classes of distributions, namely \( \alpha \)-stable, tempered stable and gamma, and for the selected distributions the estimation was done for the following parameter spaces:

- \( \alpha \)-stable: \( \alpha \in (0, 1), \sigma \in \mathbb{R}_+ \)
- tempered stable: \( \alpha \in (0, 1), \lambda \in \mathbb{R}_+, \sigma \in \mathbb{R}_+ \)
- gamma: \( k \in \mathbb{R}_+, \theta \in \mathbb{R}_+. \)

Due to the discretization of the input data the uniqueness of the solution cannot always be guaranteed. However, we have shown in our simulation study that in almost all cases the convergence to the true parameters was fulfilled. For finding the minimum distance we have used the Nelder–Mead simplex algorithm described in [50].

It is worth mentioning that the proposed methodology can be applied to broad set of distributions—in general it may serve as a method of fitting any distribution if the observations are rounded.

### 3.2. Jumps analysis

The main attention is paid to the estimation of the waiting time distribution parameters in the CTRW scenario; however, for the sake of completeness we also mention the procedure for estimating jump parameters. The problem has been widely discussed in multiple articles. In particular, in [29] it was proposed to make the probabilities of jumps dependent on a deterministic periodic function.

Due to specific properties of the real data analysed in the next section we have proposed a simplified estimation of the probabilities of jumps which follow the process

\[ X_t = \begin{cases} 
  a & \text{with probability } p_t \\
  -a & \text{with probability } 1 - p_t 
\end{cases} \]
The procedure of estimating the $p_t$ function is as follows: owing to the fact that the data analysis was based on a single day of the week (Monday) and the time series for each week were considered as realizations of the same process, $p_t$ is hence a number of positive jumps divided by the number of considered trajectories for given time point $t$. The value of parameter $a$ corresponds to the resolution of the sensor and in case of CO$_2$ concentration is equal to 50 ppm. A similar procedure was also proposed in [29] for data on indoor air quality.

### 4. Simulation study

In the simulation study we used 100 samples, each consisting of 3000 observations. The simulation scheme of the CTRW model can be found, for example, in [5]. The decision regarding the length of samples was driven by initial analysis of the data (see section 5). We assume that the jumps $X_i$ in the considered model (1) are independent identically distributed random variables equal 50 or $-50$ with probabilities 0.5. Moreover, we consider three waiting time distributions, namely $\alpha$-stable, tempered stable and gamma.

First we test our algorithm to identify the correct distribution class corresponding to the waiting times. After extraction of the waiting times from the considered trajectory we apply the algorithm to find the values of the parameters which minimize the distance in each of the distribution classes. The values of the estimators with the smallest distance indicate the distribution class that is most appropriate for data description.

The results presented in table 1 indicate that out of the three distances the most accurate in selecting the proper distribution were CvM and AD. The lower efficiency of the algorithm in the $\alpha$-stable distribution is due to low number of observations.

In the next step we have examined the robustness of our estimators with respect to various combinations of distribution parameters. The dispersion from the real parameter has been measured using mean square error; for a sample consisting of $n$ observations the error is calculated in the following way:

$$\text{MSE}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{\theta}_i - \theta)^2.$$  \hspace{1cm} (13)

In addition, we have compared our estimation procedure with the most commonly used estimators (McCulloch) in the case of the $\alpha$-stable distribution and the method of moments for tempered stable and gamma distributions.

| Table 1. Percentage of correctly identified distribution classes. Average distance is given in the bracket. |
|---------------------------------|---|---|---|
| Stable | KS | CvM | AD |
| S(0.6, 1) | 68 (0.1682) | 74 (0.0006) | 72 (0.0331) |
| S(0.75, 1) | 81 (0.0829) | 91 (0.0004) | 91 (0.0025) |
| S(0.9, 1) | 93 (0.0526) | 99 (0.0002) | 99 (0.0011) |
| Tempered stable | | | |
| TS(0.3, 1, 1) | 98 (0.0019) | 100 (0.0001) | 100 (0.0001) |
| TS(0.6, 0.5, 1) | 97 (0.0021) | 100 (0.0001) | 100 (0.0001) |
| TS(0.9, 0.1, 1) | 99 (0.0014) | 100 (0.0001) | 100 (0.0001) |
| Gamma | | | |
| G(0.5, 10) | 93 (0.0181) | 100 (0.0001) | 100 (0.0006) |
| G(2, 5) | 96 (0.0235) | 98 (0.0002) | 98 (0.0011) |
| G(5, 2) | 97 (0.0263) | 99 (0.0002) | 99 (0.0015) |
First, we examine the correctness of the proposed estimator for the $\alpha$-stable distribution. For various configurations of parameters we estimate $\alpha$ and $\sigma$. It is easy to see that the mean squared error compared with McCulloch method is significantly smaller. The performance of the method increases with higher $\alpha$ and smaller values of $\sigma$. This is obvious and stems from the fact that more observations are used for the estimation procedure; for example, for $0.6, 2\sigma$, on average for a single trajectory there were only 57 waiting times while for $0.9, 0.5$, there were 407 waiting times, see table 2.

Out of the three proposed distances CvM and AD have comparable results and in almost all cases are more efficient than KS.

For a tempered stable distribution we have examined changes with respect to $\alpha\lambda$ and $\sigma$. Since the changes with respect to $\sigma$ resemble those for an $\alpha$-stable distribution we only present the changes in terms of $\alpha$ and $\lambda$ (see table 3).

Finally we examine changes with respect to shape and scale parameters for the gamma distribution. In table 4 we compare the estimation of shape and scale parameters for proposed estimators and the method of moments. In most of the cases minimum distance estimators outperform standard methods. Out of the three proposed estimation methods the KS distance gives the best results, i.e. we obtain the lowest MSE and relatively low bias for most of the simulated waiting times. It is obvious that the estimation deteriorates when the number of observations decreases. For example, in the case of a gamma distribution with $k = 5, \theta = 10$ in a simulated trajectory of length 3000, on average only about 60 observations are used for estimation.

### Table 2. Estimation of $\alpha$-stable distribution parameters. In each cell two values are given: the upper one corresponds to the median of the estimated parameter, the lower one is mean squared error (MSE). The notation used is $S(\alpha, \sigma)$. The MSE is given in the scale $10^{-3}$.

|       | $\alpha$  |       |       | $\sigma$  |       |       |       |
|-------|----------|-------|-------|----------|-------|-------|-------|
|       | KS       | CvM   | AD    | MeC      | KS    | CvM   | AD    | MeC    |
| S(0.6, 0.5) |          |       |       |          |       |       |       |        |
|       | 0.759    | 0.626 | 0.626 | 0.650    | 0.550 | 0.535 | 0.539 | 0.733  |
|       | 26.0     | 7.0   | 7.2   | 80.2     | 42.5  | 41.9  | 24.0  | 53,689,249.6 |
| S(0.6, 1)  | 0.744    | 0.623 | 0.620 | 0.634    | 0.748 | 1.008 | 0.999 | 1.266  |
|       | 25.1     | 5.6   | 6.5   | 89.3     | 180.1 | 71.2  | 130.2 | 53,791,999.7 |
| S(0.6, 2)  | 0.775    | 0.632 | 0.628 | 0.641    | 1.083 | 1.756 | 1.778 | 2,309  |
|       | 29.6     | 7.8   | 6.7   | 147.2    | 970.4 | 444.6 | 350.2 | 100,341,231.4 |
| S(0.75, 0.5)| 0.807    | 0.752 | 0.753 | 0.746    | 0.437 | 0.498 | 0.496 | 0.526  |
|       | 3.9      | 0.8   | 0.9   | 34.6     | 7.0   | 14.9  | 3.6   | 24,657,038.0 |
| S(0.75, 1) | 0.816    | 0.752 | 0.752 | 0.765    | 0.777 | 0.992 | 0.999 | 0.989  |
|       | 6.0      | 1.5   | 1.5   | 40.0     | 77.5  | 36.9  | 31.1  | 25,456,771.7 |
| S(0.75, 2) | 0.845    | 0.763 | 0.761 | 0.762    | 1.282 | 1.893 | 1.909 | 2,153  |
|       | 10.9     | 2.8   | 2.5   | 95.9     | 653.0 | 249.6 | 171.8 | 43,178,539.9 |
| S(0.9, 0.5)| 0.918    | 0.903 | 0.901 | 0.959    | 0.480 | 0.490 | 0.493 | 0.756  |
|       | 1.3      | 0.2   | 0.2   | 37.7     | 12.4  | 5.3   | 2.5   | 24,459,941.5 |
| S(0.9, 1)  | 0.920    | 0.901 | 0.902 | 0.895    | 0.812 | 0.975 | 0.976 | 0.941  |
|       | 2.3      | 0.1   | 0.1   | 42.6     | 136.6 | 23.2  | 20.5  | 24,477,120.4 |
| S(0.9, 2)  | 0.940    | 0.902 | 0.902 | 0.900    | 1.281 | 1.926 | 1.940 | 1.960  |

$R$ Połoczański et al. J. Phys. A: Math. Theor. 50 (2017) 034002
### Table 3. Tempered stable (TS) distribution parameter estimation. The upper parameter corresponds to the median of the estimated parameter, the lower one is MSE. The notation used is TS(α, λ, σ). The MSE values are given in the scale 10^{-3}.

|     | α         |  λ         |
|-----|-----------|-----------|
|     | KS  | CvM  | AD  | MoM |
| TS(0.3, 0.1, 1) | 0.294 | 0.277 | 0.278 | 0.471 |
|     | 0.6  | 6.3  | 7.1  | 39.3 |
| TS(0.3, 0.5, 1) | 0.300 | 0.295 | 0.278 | 0.749 |
|     | 0.2  | 0.7  | 4.2  | 205.1 |
| TS(0.3, 1, 1)   | 0.318 | 0.299 | 0.342 | 0.876 |
|     | 0.7  | 0.1  | 3.1  | 331.4 |
| TS(0.6, 0.1, 1) | 0.595 | 0.692 | 0.589 | 0.644 |
|     | 0.7  | 4.3  | 4.1  | 6.2  |
| TS(0.6, 0.5, 1) | 0.598 | 0.594 | 0.530 | 0.803 |
|     | 0.3  | 1.3  | 6.4  | 42.4 |
| TS(0.6, 1, 1)   | 0.612 | 0.604 | 0.585 | 0.885 |
|     | 0.7  | 0.5  | 3.2  | 82.7 |
| TS(0.9, 0.1, 1) | 0.882 | 0.885 | 0.893 | 0.888 |
|     | 0.7  | 0.8  | 1.0  | 0.6  |
| TS(0.9, 0.5, 1) | 0.892 | 0.889 | 0.880 | 0.909 |
|     | 0.2  | 0.2  | 2.6  | 0.2  |
| TS(0.9, 1, 1)   | 0.903 | 0.897 | 0.897 | 0.930 |
|     | 0.4  | 0.3  | 0.9  | 1.0  |

### Table 4. Gamma distribution parameter estimation. The upper parameter corresponds to the median of the estimated parameter, the lower one is MSE. The notation used is G(κ, θ). The MSE values are given in a scale of 10^{-3}.

|     | κ     | θ     |
|-----|-------|-------|
|     | KS  | CvM  | AD  | MoM |
| G(0.5, 2) | 0.523 | 0.503 | 0.510 | 1.614 |
|     | 0.6  | 6.7  | 9.5  | 1266.1 |
| G(0.5, 5) | 0.517 | 0.497 | 0.505 | 0.974 |
|     | 4.9  | 6.0  | 5.0  | 230.6 |
| G(0.5, 10) | 0.516 | 0.503 | 0.502 | 0.776 |
|     | 8.4  | 8.5  | 7.4  | 85.1  |
| G(2, 2) | 1.993 | 2.016 | 2.011 | 2.180 |
|     | 2.7  | 18.9 | 17.1 | 56.3  |
| G(2, 5) | 2.061 | 2.032 | 2.038 | 2.036 |
|     | 15.2 | 19.3 | 18.9 | 40.9  |
| G(2, 10) | 2.156 | 2.101 | 2.099 | 2.001 |
|     | 22.1 | 26.2 | 26.2 | 75.3  |
| G(5, 2) | 4.990 | 5.005 | 4.950 | 4.874 |
|     | 4.7  | 31.1 | 31.6 | 235.2 |
| G(5, 5) | 5.135 | 5.144 | 4.989 | 4.766 |
|     | 18.0 | 35.3 | 36.8 | 593.0 |
| G(5, 10) | 5.200 | 5.199 | 5.199 | 4.522 |
|     | 40.0 | 39.0 | 38.6 | 1351.6 |

R Potoczański et al
5. Carbon dioxide data analysis

In the following section we will perform analysis of CO₂ data. In section 5.1 we describe the data and introduce the data cleansing procedure. In section 5.2 we model the data using the CTRW process with the main emphasis on waiting time analysis.

5.1. Data description

Measurements of CO₂ concentration were performed in a lecture room with an amphitheatre-like layout. Room dimensions were 19 × 8 × (4–2.9 m). The room had only one external wall, fitted with six huge openable windows. Despite the availability of mechanical ventilation, air exchange is realized predominantly via natural ventilation. Teaching hours extend from 9:00 to 21:00. Classes are held during all working days and on the majority of weekends (part time studies). Teaching blocks are typically 1.5 h long with the breaks of 15 min in between. Although designed for 90 students, the lecture room is hardly ever occupied to that extent. In the examined period, audience numbers changed considerably within a single day as well as on a day to day basis.

CO₂ measurements were performed in the central part of the room at a height of about 1 m. The measuring device was separated from the direct influence of the emission sources (students and the teacher). The monitoring was realized with the instrument dedicated to continuous measurements and data logging. It is based on the NDIR sensor and offers a measuring range of 0–5000 ppm (accuracy 50 ppm +3% of measured value and the measurement data resolution 1 ppm). This level of performance may be currently considered as a standard in indoor air quality studies.

The measurement results were recorded with a time resolution of 15 s. The data was collected during nine consecutive weeks, between 20 February and 22 April 2013. Since the lecture hours are different for each day, in this paper we focus only on Mondays—the analysis

![Figure 5. Carbon dioxide concentration on eight consecutive Mondays from 25 February until 22 April. Due to holidays, no lectures were held on 1 April.](image)
can be easily replicated for other weekdays. The trajectories of eight consecutive Mondays are displayed in figure 5.

By looking at the chart it is easy to notice that during lecture hours the CO$_2$ concentration rises significantly and decreases during the breaks. However, in contrast to the analysis of temperature data presented in [24], no particular trend can be observed in the data. Additionally, due to irregular student attendance at lectures, estimation and prediction of CO$_2$ concentration based on the number of students subscribed to the lectures might be inaccurate.

Taking into account teaching hours and the measurement device accuracy we have performed ‘cleansing’ of the data in order to reveal only significant changes in the data:

- Lectures were only held until 20:00. By about 21:00 the process stabilized around a certain level and only small fluctuation can be observed until the next day’s lectures. We called the time from 21:00 to 09:00 the stale period and removed it from core analysis. As estimation of the stale period is of interest, and the average value fluctuates around 400 ppm, so we assume that during the stale period the CO$_2$ concentration is equal to 400 ppm.
- The device accuracy is 50 ppm ±3% of the measured value. A difference of such magnitude indicates the change in the state of the indoor air with respect to CO$_2$ concentration, which is greater than the measurement error. Hence, it should be respected. For this reason we have rounded the observed values to the closest multiple of 50.
- Furthermore, we have removed periods where the value has changed for a short period—less than 30 s—and then returned to the previous level.

5.2. Data analysis

After applying the cleansing procedure described in section 5.1, we considered CTRW as a stochastic system that allows data modelling. At this point let us also comment on the important issue of aging of various physical systems [52, 53]. This phenomenon is of great importance in data analysis, since the moment when we start to collect (measure) the data significantly
influences, for instance, the distribution of waiting times. Here we assume that the evolution time of the system is very long compared with the time when we started to measure the data. Thus waiting times can be reasonably described by the usual theoretical distributions like $\alpha$-stable, tempered stable or gamma.

In the first stage we have divided the analysed data into two vectors: the first one corresponding to waiting times and the second consisting of jump sizes. The vectors representing the waiting times are presented in figure 6 while the corresponding jumps are displayed in Figure 7.

Figure 7. The vectors of jumps on eight consecutive Mondays from 25 February to 22 April.

Figure 8. The sample autocorrelation function of constant time periods series (left panel) and corresponding squared series (right panel) for 25 February.
We have assumed that the data constitutes eight realizations of the same stochastic process. Each sample has been analysed independently in the context of waiting times and probability of upward–downward movement.

In order to make use of the CTRW to model the data, it needs to be validated if the vector \( \{ T_i \} \) of waiting times forms an independent, identically distributed sample. For each of eight samples we test the independence with a visual check of the autocorrelation function of the series and the squared series as described in [35]. In figure 8 we present the obtained autocorrelation functions for the first sample (25 February). As can be observed, the calculated values are close to 0 and most of them lie within the 95% confidence intervals. The procedure has been applied to all eight samples and the results resemble those presented in figure 8. Hence, we may conclude that the waiting times are independent. The identity of distributions can be verified, for example, by testing the behaviour of the empirical second moment of the time series as described in [51].

The estimation of parameters has been performed using the method described in section 3. As shown in our simulation study in section 4, the best fit was obtained using the AD distance,
hence this criterion was used for parameter estimation. In figure 9 we present a fit of the rescaled modified CDF for all three distributions—\(\alpha\)-stable, tempered stable and gamma—to the empirical distribution function on 25 February. As shown in the plot, the gamma distribution provides a relatively good fit to the data and outperforms the other distributions.

Out of eight samples, in 62.5\% of cases the best fit was obtained using the gamma distribution, in 25\% cases tempered stable was selected and in 12.5\% cases \(\alpha\)-stable. Furthermore, the differences between the gamma distribution and the chosen ones in these 37.5\% of cases was negligible. This indicated that the gamma distribution is the most adequate for a description of waiting times.

The results of fitting the gamma distribution are shown in table 5: the estimated values of the shape parameter are similar for each sample but there is more variability in the estimated values of scale parameter. For further processing and simulation we used the median of both parameters.

Figure 10. Probability of upward movement during lecture hours.

Figure 11. \(\text{CO}_2\) concentration on 25 March (Monday) during lecture hours together with simulated quantile lines on levels 0.1, 0.5 and 0.9.
The probability of upward and downward movements is highly dependent on several factors, in particular whether lectures are held at that time and how many students attended the lecture. Due to variability of these factors, fitting of any trend function is problematic and the results might be unreliable. Taking this into account, we have decided to model the probability based on the hour of the day; the procedure is described in section 3.2. The result of the fit is presented in figure 10.

Taking into account the above-mentioned facts we have used the estimated parameters of waiting time distribution and the probability of upward and downward movement to simulate the process. In the period marked as stale—from 09:00 to 21:00—the process stays at the same level (400 ppm). During the lecture hours the process follows a CTRW with waiting times from the gamma distribution with a shape equal to 0.4188 and scale equal to 21.5421. In addition we have applied constraints such that after 21:00 the process tends to the stale value—one it is reached it stays at this level. We have simulated 10000 trajectories of the CTRW model and matched it with our data samples: as an example a comparison of a sample from 25 March with quantile lines on levels 0.1, 0.5 and 0.9 is shown in figure 11. We conclude that the simulated process gives a relatively good fit to the data and could be used as a predictor of future values. We recall that in statistics and the theory of probability, quantiles are cutpoints dividing the range of a probability distribution into contiguous intervals with equal probabilities, or dividing the observations in a sample in the same way. The quantile \( x \) of order \( p \) for a continuous random variable \( X \) satisfies

\[
P(X \leq x) = p.
\]

On the basis of the above inequality, the empirical quantiles can be calculated for IID observations being realizations of the random variable \( X \).

Major practical implications of the proposed approach may be demonstrated in the domain of continuous measurements. This consists of the precise estimation of the waiting times. Such estimation allows us to determine the distribution of time until the process changes significantly. From the point of view of measurement it is the guideline for how frequently the state of the measured parameter should be checked without losing relevant information about its temporal variation. More specifically, the distribution mean may be considered as an indication of the optimum sampling frequency.

In table 5 we present the parameters of the most suitable distributions for the waiting times which were fitted to the CO2 concentration data collected during eight consecutive Mondays in the lecture room. As shown, the mean values of the distribution were from 1 min 45 s to 2 min 30 s. As an example, on the day with the minimum mean, we see that most frequently meaningful changes of CO2 concentration occurred every 1 min 45 s. Consequently, if the measurement of the parameter is performed less frequently, significant information may be lost. The inverse of the quoted time interval could be considered as the suggested sampling frequency. A smaller frequency would not be recommended. Considering that the analysed data were collected every 15 s, the obtained result of 1 min 45 s indicates the possibility of saving considerable data storage space. By adjusting sampling frequency, the number of data collected is seven times smaller. Hence, the duration of continuous measurements between consequent data download may be extended seven times. This result is of high practical value. However, we should emphasize that it does not have the status of a general recommendation. It is valid for the CO2 concentration measurements performed with a defined accuracy of 50 ppm in this particular indoor environment. More studies need to be performed in order to obtain more generic results applicable in measurement practice in different circumstances. Because the methodology can be applied to other parameters, e.g. temperature or relative humidity,
the approach presented in this paper bears considerable consequences for planning indoor air quality monitoring networks, in particular the use of portable devices.

6. Conclusions

In this paper we have considered a stochastic system that allows modelling of the time series of CO₂ concentration that is one of the main parameters of indoor air quality. The model we have applied for data analysis is the CTRW, which exhibits anomalous diffusion behaviour. The main goal of our work was to propose new techniques for estimation and statistical investigation of the distribution of waiting times. These methods are based on the modified CDF, an extended version of the classical CDF, which is more appropriate for rounded data such as waiting times visible in CTRW trajectories. By using simulated data we have proved the efficiency of the proposed techniques. For investigation of the robustness of our method we have considered three waiting time distributions commonly used in practice, namely α-stable, tempered stable and gamma. Moreover, we have supported theoretical and simulated results with real data analysis.

The approach was applied to analyse time series of CO₂ concentration data recorded in an indoor environment. It was demonstrated that the method leads to valuable conclusions concerning sampling frequency. This is one of the crucial factors that has to be taken into account while planning continuous measurements, although most frequently its best value is not known. The presented approach is general enough to be applied to almost any measured parameter. In addition the method may be useful in determining sampling frequency in diverse problems involving continuous measurements.

Acknowledgments

This paper is a result of the project ’The variability of physical and chemical parameters in time as the source of comprehensive information about indoor air quality’. The research of MM, AS and AW is co-financed by the National Science Center, Poland, under the contract No. UMO-2012/07/B/ST8/03031. The research of JG was partially supported by NCN Maestro Grant no. 2012/06/A/ST1/00258.

References

[1] Meroz Y and Sokolov I M 2015 Phys. Rep. 573 1
[2] Montroll E W and Weiss G H 1965 J. Math. Phys. 6 167
[3] Metzler R and Klafter J 2000 Phys. Rep. 339 1
[4] Magdziarz M 2009 Stoch. Process. Appl. 119 3416
[5] Magdziarz M, Weron A and Weron K 2007 Phys. Rev. E 75 016708
[6] Magdziarz M 2010 Stoch. Models 26 256
[7] Piryatinska A, Saichev A I and Woyczynski W A 2005 Physica A 349 375
[8] Metzler R, Jeon J-H, Cherstvy A G and Barkai E 2014 Phys. Chem. Chem. Phys. 16 24128
[9] Hirling F and Franosch T 2013 Rep. Prog. Phys. 76 046602
[10] Metzler R, Jeon J-H and Cherstvy A G 2016 Biochim. Biophys. Acta—Biomembr. 1858 2451
[11] Scher H and Montroll E 1975 Phys. Rev. B 12 2455
[12] Scher H and Lax M 1973 Theory. Phys. Rev. B 7 4491
[13] Pfister G and Scher H 1978 Adv. Phys. 27 747
[14] Ott A, Bouchaud J-P, Langevin D and Urbach W 1990 Phys. Rev. Lett. 65 2201
[15] Caspi A, Granek R and Elbaum M 2000 Phys. Rev. Lett. 85 5655
[16] Golding I and Cox E C 2006 Phys. Rev. Lett. 96 098102
[17] Bronstein I, Israel Y, Kepten E, Mai S, Shav-Tal Y, Barkai E and Garini Y 2009 Phys. Rev. Lett. 103 018102
[18] Hu X, Hong L, Smith M D, Neusius T, Cheng X and Smith J C 2016 Nat. Phys. 12 171
[19] Ali Tabei S M, Burov S, Kim H Y, Kuznetsov A, Huynh T, Jureller J, Philipson L H, Dinner A R and Schere N F 2013 Proc. Natl Acad. Sci. USA 110 4911
[20] Weigel A V, Simon B, Tamkun M M and Krapf D 2011 Proc. Natl Acad. Sci. USA 108 6438
[21] Orzel S and Wylomańska A 2011 J. Stat. Phys. 143 447
[22] Janczura J, Orzel S and Wylomańska A 2011 Physica A 390 4379
[23] Maciejewska M, Szczurek A, Sikora G and Wylomańska A 2012 Phys. Rev. E 86 031128
[24] Janczura J, Maciejewska M, Szczurek A and Wylomańska A 2013 J. Stat. Phys. 152 979
[25] Szczurek A, Maciejewska M, Teuerle M and Wylomańska A 2015 Physica A 420 190
[26] Gajda J and Magdziarz M 2010 Phys. Rev. E 82 011117
[27] Stanislavsky A, Weron K and Weron A 2008 Phys. Rev. E 78 051106
[28] Stanislavsky A, Weron K and Weron A 2014 J. Chem. Phys. 140 054113
[29] Janczura J and Wylomańska A 2012 Acta Phys. Pol. B 43 1001
[30] Ramirez-Nino J, Pascacio A, Carrillo J and de la Torre O 2009 Measurement 42 1203
[31] Ferreira L F, Antunes P, Domingues E, Silva P A and Andr P S 2012 Measurement 45 1527
[32] Zhao L, Xie Y, Wang J and Xu X 2015 Atmos. Environ. 122 382
[33] Martinez-Garrido M J and Fort R 2016 Measurement 82 188–201
[34] Govindarajulu Z 1999 Elements of Sampling Theory and Methods (Englewood Cliffs, NJ: Prentice Hall)
[35] Wylomańska A 2012 Physica A 391 5685
[36] Neumann M H and Reiss M 2009 Bernoulli 15 223
[37] Klafter J and Sokolov I M 2011 First Steps in Random Walks: from Tools to Applications (Oxford: Oxford University Press)

References