Measurement 46 (2013) 727–738

Contents lists available at SciVerse ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Modeling of probability distribution functions for automatic threshold calculation in condition monitoring systems

A. Jablonski [a,*], T. Barszcz [a], M. Bielecka [b], P. Breuhaus [c]

[a] AGH University of Science and Technology, Faculty of Mechanical Engineering and Robotics, Mickiewicza 30, 30-059 Kraków, Poland
[b] AGH University of Science and Technology, Faculty of Geology, Geophysics and Environmental Protection, Mickiewicza 30, 30-059 Kraków, Poland
[c] International Research Institute of Stavanger AS, Postboks 8046, 4068 Stavanger, Norway

ARTICLE INFO

Article history:
Received 27 April 2012
Received in revised form 22 August 2012
Accepted 25 September 2012
Available online 17 October 2012

Keywords:
Condition monitoring
Automated threshold setting
Probability distribution
False alarms

ABSTRACT

A suitable setting of threshold levels has been a dilemma for engineers in a number of science and industrial fields. It is a common problem in monitoring, where deviation from a correct state needs to be detected. As the number of monitored values in modern systems reaches thousands, threshold calculations became a significant yet frequently underestimated concern. Due to a prohibitive cost of manual threshold setting, many systems generate thousands of false alarms consequent upon running on default threshold levels. In the paper, the authors illustrate a methodology for automatic threshold calculations in a large monitoring system. The paper is mainly addressed to engineers and machine monitoring systems developers, therefore selected statistical topics were treated briefly with main focus on practical solutions. Two fundamental data types are considered; namely, *vibration signal measures*, which can be extended to any nonnegative data, and symmetrical *process values*. As it is shown, these data types have significantly different probability distributions. Since real data seldom fits the Gaussian model, an investigation of several distributions and their comparison is presented. The proposed approach is validated on four datasets including process values from a gas compressor and vibration signal measures from a wind turbine.

1. Introduction

Modern systems of monitoring and diagnosis (SM and D) experience a continuous growth in a number of aspects. Firstly, these systems take advantage of a growing computational power of available computer hardware along with more reliable software and network solutions. Secondly, they become widespread due to their ultimate economical benefit, which frequently compels the integration of a monitoring system early at the machine designing stage. Currently, the wind power generation might be considered as one of globally fastest growing markets. The industrial practice of the authors concerning the data processing for threshold settings for such systems has led them to certain observations and conclusion, which are stated in the paper.

Threshold configuration of digital data sets became a challenge in large distributed data acquisition and condition monitoring systems. Generally, the manufacturers of monitoring and diagnostic systems construct their products on the basis of official guidelines, for instance available in the insurance companies documentation [1–3]. These guidelines require the generation of numerous alarm and warning threshold for all calculated characteristic frequencies of mechanical components. Threshold levels are next reproduced for all defined operational states of a machine. Consequently, given the number of characteristic frequencies multiplied by the number of states and by the number of vibration sensors results in hundreds of monitored values per one turbine, and in thousand monitored values per a typical wind farm [4]. One solution of

[*] Corresponding author.
E-mail addresses: ajab@agh.edu.pl (A. Jablonski), bielecka@agh.edu.pl (M. Bielecka), peter.breuhaus@iris.no (P. Breuhaus).

0263-2241/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.measurement.2012.09.011

threshold settings is to perform it manually. However, taking into account the above, it is clear that manual settings require time consumption (and a corresponding financial burden) far greater than allowable in practice.

Moreover, from the ergonomic point of view, as time passes, a user sets alarm levels with less caution. Unfortunately, many times a setting alarm level is a tradeoff between missing an alarm and generating false alarms. Setting an alarm level without sufficient attention on one hand might generate a huge number of false alarms for a level set too low, while on the other hand it might not detect the state of malfunction for a level set too high.

Surprisingly, the number of references dealing with the subject is relatively low, mainly due to commercial confidentiality clauses. There are several works, which analyze the problem with the assumption of the normal distribution of the data [5]. There were also works of Cempel [6,7], who proposed a method to calculate so called "limit values" by statistic methods. Cempel was analyzing the data during extended period of time (so called "passive experiment"), from which the limit values were based on the estimation of the mean life curves. This method turned out to be an important contribution to the problem of prognostics.

Recently, numerous methods of fault prognostics were presented in a review paper by Heng et al. [8], accompanied by a large number of references. In Ref. [9], the algorithm is based on the density of random occurrences, which not only is applicable to a narrow field of gearboxes, but also is prone to external signal randomness, which is inevitably mixed with the anticipated randomness reflecting the degree of a mechanical element's deterioration. In [10], the proposed algorithm is constrained by extra information usually unavailable in large monitoring system, which implies its laboratorial (e.g. test rig) character. Finally, in [11] the authors move directly to the issue of a general management after alarm's occurrence. Nevertheless, none of these contributions have addressed the issue of automatic threshold level setting. Thus, there is a crucial need for the development of automatic procedures for threshold settings in large databases.

The initial attempts to set limit values without using model fitting have shown strong variations between data sets from the same parameter monitored in different time period, even after selecting data from a vicinity of a given machine operating point. These results have encouraged the authors to investigate the application of probability distribution function (PDF) modeling, which turned out to be able to minimize the differences between data from various time periods.

Since the paper illustrates some data from wind turbines working under non-stationary operating conditions, it is important to comment that the analyzed data has been filtered according to predefined machine operational states. Consequently, this "state constraint" requires continuous consideration of process parameters during acquisition process, and removal of data corresponding to transient states [3]. On the other hand, as illustrated by Bartelmus and Zimroz in [12,13] and Zimroz in [14], signals recorded during significant change of operating conditions might be analyzed as well. In the latter case, the consideration of process parameters takes place not during acquisition but during reasoning process. Nevertheless, regardless of selected approach, it is crucial to take into account "different amplitudes and properties" of vibration signals characteristic for variable operational parameters [12].

The paper is organized as follows: Section 2 describes requirements of a general method for threshold settings and gives additional insight to data handling. Section 3 presents a basic mathematical background and reviews selected PDFs, which are considered to be potentially useful for machine data fitting. Section 4 presents the data investigated in the paper. Section 5 describes the proposed method for threshold settings. Two approaches for threshold settings named "symmetrical" and "positive amplitude" are illustrated. A flowchart of the method is illustrated with a detailed description of each level, including both mathematical and practical considerations. Section 6 illustrates the validation of proposed method on real data from a gas compressor and a wind turbine. The results of statistical fitting for all investigated distributions are presented. The reader should be aware that the paper is engineering oriented with emphasis on practical solutions, and therefore the mathematical apparatus describing statistical methods used throughout the paper is adequately simplified.

2. Threshold setting requirements

Intuitively, a threshold setting scheme consists of the data and the rules. The data generally is arranged as two-column vectors containing amplitudes with respect to timestamps. The rules refer to a set of consecutive mathematical operations, which process the data, and produce threshold levels. Few remarks about each element are worth mentioning.

The *data* contains: (1) true information about the process, and (2) true data with amplitude deviations. These deviations are mainly due to:

(a) Transducers imperfections.
(b) Changing operational conditions.
(c) Random peaks of various external source.
(d) Electronic failure signals.
(e) Machine failure signals.

As it will be shown in the paper, application of a model to the raw data increases the chances to filter the true process behavior and to minimize the influence of listed measurement inaccuracies.

The *rules* applied to the real data constitute the critical part of the system. Without an application of a general algorithm, writing rules to be applied to a raw data would be a painstaking or even impossible task. Application of an algorithm enables implementation of a set of rules to data series without the need for extra data handling. If a procedure for threshold settings is to be applied automatically, it is obliged to fulfill strict requirements.

Firstly, such procedure needs to be feasible for signals from various machines. Interestingly, even among wind turbines from different manufacturers, various kinematic designs might actually be considered as separate machine types due to major differences in signal characteristics.

Secondly, the procedure ought to be flexible to data of a wide range of amplitude orders. In case of vibration data measures, amplitudes vary from relatively small envelope-based signatures (e.g. rolling element bearings fault frequencies) to large time signal-based scalar estimators (e.g. peak-to-peak value).

Tertiary, the input data to the algorithm needs to be filtered according to defined machine's operational points (often called "states"). Moreover, the algorithm has to be invulnerable against random data amplitudes frequently present in a vibration data measures. Customarily, threshold settings are set by means of a warning level, which corresponds to a significant difference in amplitude, and an alarm level, which indicates a severe change in amplitude level. Once the optimal levels for both alert and alarm are set, the procedure needs to assume an acceptable degree of error, which is a separate task, and depends on the machine major failure acceptance. A suitable acceptance level might be found with the aid of true and false rates classifying tools, for instance receiver operating characteristics (ROCs) curves or it is more recent area under the curves (AUCs) estimators [15,16].

3. Probability distributions for threshold setting

While working in the field of industrial monitoring systems development, the authors have noted that the most common approach to the subject of threshold levels is to apply a Gaussian distribution to raw vibration data measures. In such case, the natural solution is to calculate the average value μ and the standard deviation σ of a data set (e.g. peak-to-peak trend). Then, the theoretical probability P that the data from the data set will fall within the range $\pm k\sigma$ (k is an integer) is:

$$P(|X-\mu| < k\,\sigma) = \begin{cases} 0.6827 & \text{for } k=1, \\ 0.9545 & \text{for } k=2, \\ 0.9973 & \text{for } k=3. \end{cases} \tag{1}$$

The main advantage of the method is its simplicity, but this assumption is valid only for the normal distribution. Nevertheless, real data is hardly ever "nicely" normally distributed, which results in an enormous number of false alarms generated for such approach or very high values of the coefficient k, and high number of missed alarms.

In case when data does not fit to the normal distribution, intuitively, it is appropriate to find the distribution, which is close enough to model the data. After initial research, the authors have decided to focus attention on four distributions: Weibull, generalized extreme value (GEV), extreme value and inverse Gaussian. Sections 3.1–3.4 give a brief overview of each distribution.

3.1. Weibull probability distribution

A random variable x is said to have a Weibull probability distribution with parameters a and b if its density function is given by following formula [17]:

$$\begin{aligned} f_{a,b}(x) &= ab^{-a}x^{a-1}\exp\left[-\left(\frac{x}{b}\right)^{a}\right]1_{(0,\infty)}(x) \\ &= \frac{a}{b}\left(\frac{x}{b}\right)^{a-1}\exp\left[-\left(\frac{x}{b}\right)^{a}\right]1_{(0,\infty)}(x), \end{aligned} \tag{2}$$

where a is the shape parameter, b is the scale parameter. The Weibull distribution is one of the most widely used lifetime distribution in reliability engineering. It gives the distribution of failures, where the failure rate is proportional to a power of time. As stated by Cempel in [7], Weibull distribution should be used for object with uniform wear. Graphs of Weibull distribution for various values of parameters a and for $b = 1$ are presented in Fig. 1.

Although various numerical and graphical methods for estimating parameters of the Weibull probability distribution function from the data set are available, the most frequently used are methods of moments, the Maximum Likelihood method, and the least square method. However, the Maximum Likelihood method [18] has proved to be the most efficient.

3.2. Generalized extreme value probability distribution

The probability density function for the generalized extreme value distribution with a location parameter a, a scale parameter b, and a shape parameter $k \neq 0$ is given as

$$\begin{aligned} f_{a,b,k}(x) = b^{-1}\exp\left(-\left(1+k\cdot\frac{(x-a)}{b}\right)^{\frac{-1}{k}}\right) \\ \cdot\left(1+k\cdot\frac{(x-a)}{b}\right)^{-1-\frac{1}{k}}. \end{aligned} \tag{3}$$

For

$$1+k\cdot\frac{(x-a)}{b} > 0, \tag{4}$$

the distribution corresponds to the Type II case, while $k < 0$ corresponds to the Type III case. In the limit for $k = 0$, corresponding to the Type I case, the density is given by following formula:

$$f_{a,b,0}(x) = b^{-1}\exp\left(-\exp\left(-\frac{(x-a)}{b}\right)-\frac{(x-a)}{b}\right). \tag{5}$$

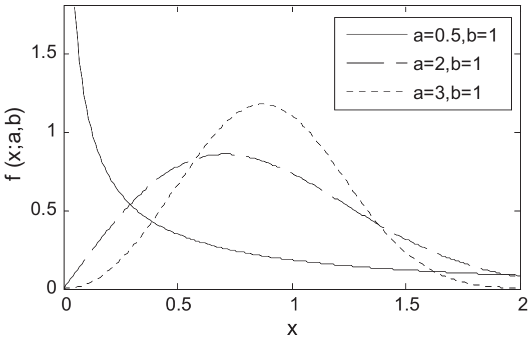

Fig. 1. Weibull probability distribution function for various values of the parameter a and for parameter $b = 1$.

Graphs of the generalized extreme value distribution for various value of parameters a, b and k are presented in Fig. 2.

The generalized extreme value distribution is often used to model the smallest or largest value among a large set of independent, identically distributed random values representing measurements or observations. Therefore, it seems reasonable to consider it as a potential candidate for threshold setting.

3.3. Extreme value probability distribution

A random variable x is said to have an extreme value probability distribution with a location parameter a and a scale parameter b if its density function is given by following formula:

$$f_{a,b}(x) = b^{-1}\exp\left(\frac{x-a}{b}\right)\exp\left(-\exp\left(\frac{x-a}{b}\right)\right). \tag{6}$$

Graphs of the extreme value distribution for various values of parameters a and b are presented in Fig. 3.

The extreme value distribution is appropriate for modeling the smallest value from a distribution whose tails decay exponentially fast, for example, the normal distribution.

3.4. Inverse Gaussian probability distribution

A random variable x is said to have an inverse Gaussian probability distribution with parameters $a > 0$ and $b \geqslant 0$ if its density function is given by following formula:

$$f_{a,b}(x) = \frac{a}{\sqrt{2\pi}}x^{-3/2}\exp\left(-\frac{(bx-a)^2}{2x}\right). \tag{7}$$

Graphs of the inverse Gaussian distribution for various values of parameters a and b are presented in Fig. 4.

Also known as the Wald distribution, the inverse Gaussian is used to model nonnegative positively skewed data. The distribution originated in the theory of Brownian motion, but has been used to model diverse phenomena. Inverse Gaussian distributions have many similarities to standard Gaussian (normal) distributions, which lead to applications in inferential statistics.

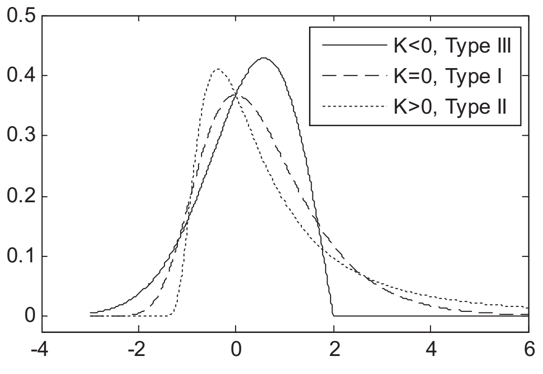

Fig. 2. GEV probability distribution function for various values of the parameters a and b [18].

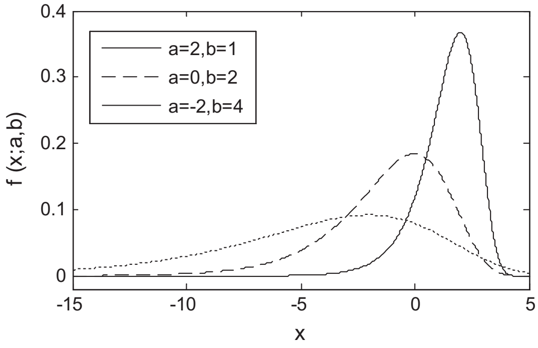

Fig. 3. Extreme value probability distribution function for various values of parameters a and b [19].

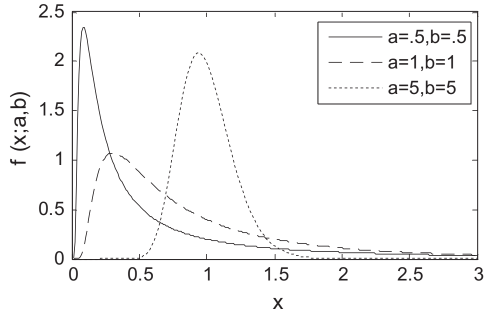

Fig. 4. Inverse Gaussian probability distribution function for various values of the parameters a and b.

4. Description of real data

The chapter presents a set of four exemplary data trends from various sources of different types.

- High amplitude vibration-based estimator (peak-to-peak trend value from a gas compressor), referred to as set *A*.
- Low amplitude vibration envelope-based estimator (energy of the outer ring band of a rolling element bearing from a gas compressor), referred to as set *B*.
- High amplitude process channel (temperature from a gas compressor), referred to as set *C*.
- Low amplitude vibration envelope-based estimator (energy of the outer ring band of a rolling element bearing from a wind turbine), referred to as set *D*.

The first data set is typical for high amplitude direct signal estimator. The second and the fourth data sets are typical for the energy values from a defined narrow frequency range, i.e. for post-processed data.[1] The first, second, and

[1] Sets *B* and *D* refer to envelope-based measures calculated as follows: a recorded vibration signal is classified according to states definition. The signal is high-pass filtered (1 kHz), rectified, and low-pass filtered (500 Hz). Next, a signal FFT is calculated, and a spectrum fragment corresponding to a particular machine component characteristic frequency and its neighborhood (e.g. 2%) is selected. Finally, signal energy from selected spectrum fragment is calculated as a single scalar value.

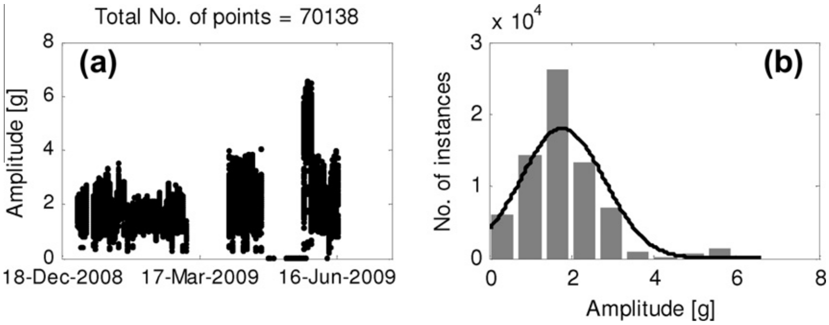

Fig. 5. Peak-to-peak trend value from a gas compressor (a) data trend, and (b) data histogram with (normalized) Gaussian PDF fit.

fourth data sets were collected at a predefined machine operational state selected according to speed and power (in case of wind turbine) measurements, and are examples of one-sided, positive data. The third data set is typical for symmetrical process parameters data. The first set has relatively high energy compared with the second and fourth set, which causes different probability distributions.

Sets B and D both represent data calculated form fragments of spectra of demodulated signals. However, set B represents data, which corresponds to bearing minor defect, and which was replaced at an incipient fault stage (note 4 distinct phases of the data, i.e. steady, rise, replacement followed by high vibration levels due to initial exploitation stage, and steady phase again). On the other hand, set D shows data free from any bearing fault symptoms. As mentioned before, these values were calculated form envelope spectrum; therefore, in case of set D they represent variations of noise levels rather than fluctuations of vibration levels. Consequently, its PDF (both filtered and unfiltered sets) is clearly different from PDFs of other data sets.

Figs. 5–8 illustrate the selected data. In each case, the figure on the left side shows the entire data amplitudes against time, and the figure on the right side shows the data histogram (with ten bins) together with an optimal fit[2] of a Gaussian probability distribution function.[3]

From Figs. 5–8 it is clear that a normal distribution does not fit any of presented data accurately. It might be argued that Figs. 5 and 6 represent a normally distributed data, but statistical tests for normality (for instance Jarque–Bera test for with 5% significance level) prove the assumption wrong. This fact is of utmost importance especially in monitoring systems installed on machines with varying operational conditions, which is the case for wind turbines. The application of operational states does not eliminate the undesired data spread neither. Therefore, an alternative (more robust) approach for data modeling needs to be sought. Following chapters present authors' proposals in this matter.

5. Threshold setting procedure

In Section 4, two data types were introduced, a vibration-based estimators, and a process channel values. The authors believe that these two groups actually constitute a top-level division of signals in monitoring systems in terms of threshold procedures to be applied, because these two groups differ greatly in signal characteristics. Consequently, two distinctive approaches for threshold settings are presented. These approaches are relatively simple; yet, they proved themselves to be a significant observation.

5.1. Symmetrical approach

This approach is applicable for process data, the intrinsic behavior of which tends to be symmetrical, i.e. is inclined to oscillate around a mean value. Such data includes for instance temperature or pressure. In this group, data might accept both positive, negative, and zero values. It is important to realize the possibility of non-positive values, since some mathematical formulas do not accept them; therefore, the threshold algorithm might require extra programming effort. It is a common practice that in this case, the reference threshold levels need to be set on lower and upper boundary levels. Fig. 9 illustrates how the *Warning* and *Alarm* levels might be located for this type of data.

Thus, the main task in the threshold calculations requires analysis of data variability in both directions, i.e. towards minimum and towards maximum.

5.2. Positive amplitude approach

This approach is designed for parameters calculated for instance from the vibration data measures (in vast majority of cases energy of a peak value in a band from a spectrum), which by definition is non-negative. Vibration data measures are usually characterized by low mean level with a relatively large number of peaks. For modeling purposes, it might be assumed that the nature of data is random with an unknown distribution.

[2] Fits were conducted with the Matlab® software.

[3] Note that the figures on the right side are dimensionless. It is because on one hand the height of the bins is a positive integer number of occurrences, while on the other hand the total are under the PDF curve converges to one. Therefore, if these two graphs are to be plotted on a single axis, one figure must be scaled. Since the fits are only for visual assessment, the scales were omitted in order to avoid their ambiguity.

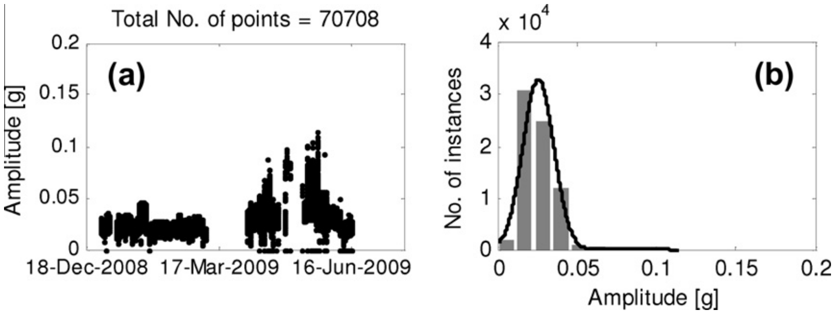

Fig. 6. Outer ring of a rolling element bearing from a gas compressor (a) data trend, and (b) data histogram with (normalized) Gaussian PDF fit.

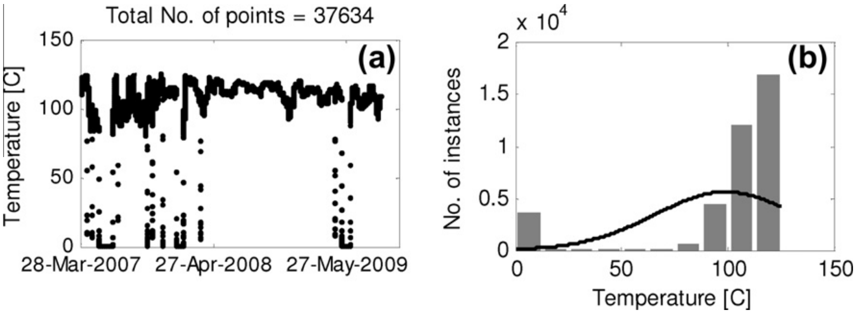

Fig. 7. Temperature from a gas compressor (a) data trend, and (b) data histogram with (normalized) Gaussian PDF fit.

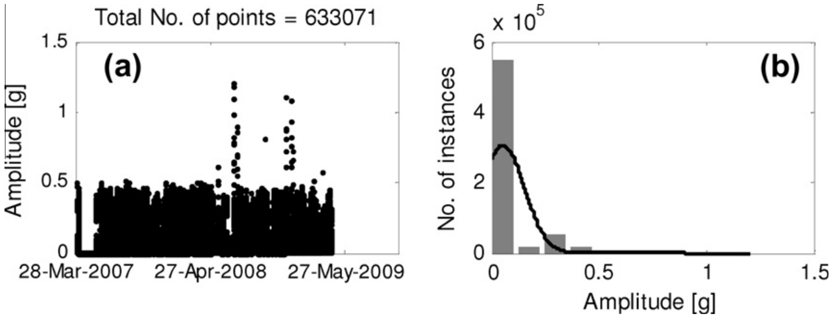

Fig. 8. Rolling element bearing from a wind turbine (a) data trend, and (b) data histogram with (normalized) Gaussian PDF fit.

However, for the majority of a real, industrial data, the median value is usually much smaller than the mean value, which disqualifies the usage of the normal distribution-based estimators. Fig. 10 illustrates how the *Warning* and *Alarm* levels might be located for this type of data. Thus, the main task becomes finding the limit only in the positive direction. The two types of data can be processed with the same algorithm, however sometimes it is possible that different distributions will yield different fits to the data. This problem is dealt with later in the paper.

5.3. The method flowchart

The proposed method consists of a combination of several steps, as presented in Fig. 11.

The first three steps describe data preprocessing phase. The fourth one is the core of the method. In the last two, thresholds are calculated on the basis of distributions.

5.4. Data preprocessing

5.4.1. States definition

Due to various contents of data (as mentioned in Section 2), data needs to be preprocessed (i.e. filtered) before it is modeled. In a majority of commercial systems of monitoring and diagnosis, the data preprocessing is performed explicitly by machine operational states definition. However, this technique is not flawless, since in reality every sample is registered in a unique operational state. Therefore, states are usually defined as ranges of defined process estimators. As a consequence, the randomness in the

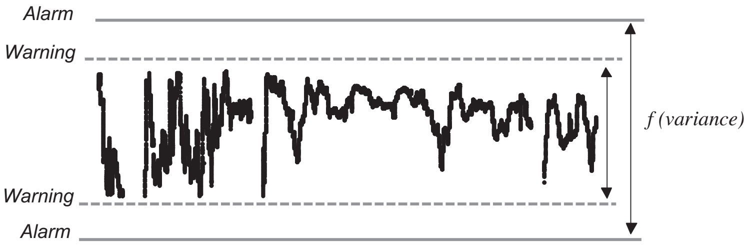

Fig. 9. Illustration of *Warning* and *Alarm* levels as a function of variance.

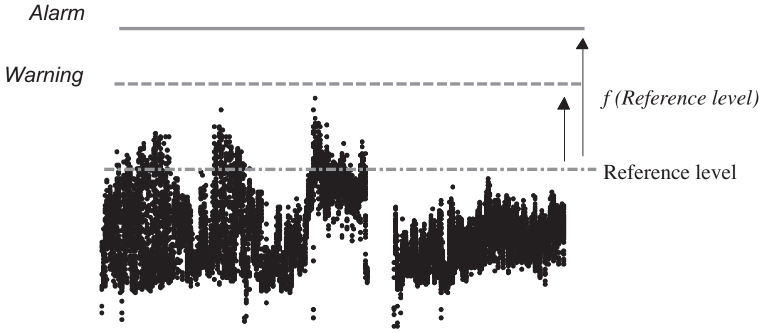

Fig. 10. Illustration of *Warning* and *Alarm* levels as a function of reference level.

state-defined data seems to be inevitable (which is endorsed by the authors' industrial practice).

5.4.2. Outliers removal

Outliers are understood as values registered "by a chance", which do not represent a true process behavior, but might significantly influence conclusions about the data. The outlier removal is an extensive field of statistical data analysis and computer science. Although numerous theories on outlier detection exist [20,21], criteria for an "outlier" have not been explicitly defined so far. Therefore, the outlier removal is virtually a subjective issue. Outliers' detection is most frequently based under the assumption of normally distributed data. In this case, the most popular detection methods include: Peirce's criterion [22], Grubbs' test (Maximum Normed Residual Test) [23], the Generalized ESD (Extreme Studentized Deviate) [24], Chauvenet's criterion [25], and Dixon's Q test [26]. It is important to point out that all but last methods involve calculation of the mean value, standard deviation, and a rigid definition of the significance level (usually denoted as alpha). Moreover, some of the tests assume a known number of outlier points in a given data set (for instance a single value for a Dixon Q test). Because of these constrains, aforementioned tests are not feasible for data considered in the paper. Firstly, the significance level is always an individual choice itself, and cannot be straightforwardly calculated. Secondly, the number of outliers in real data is unknown. The authors have conducted a research on a more universal vibration signal measure outlier detection method, which would overcome distribution-related and quantitative limitations. In terms of machine diagnostics, outliers refer to trend data, which might be of two types, namely, process values and vibration signal measures. While the former group refers to a raw signal measurement (temperature, pressure, etc.), the latter is associated with a signal energy (or at least its peak-to-peak value). Therefore, trend data values (as well as outliers) are susceptible to machine operational conditions, especially close to state's limits or to local resonances, through which a given machine passes. Furthermore, outliers are frequently result of invalid data

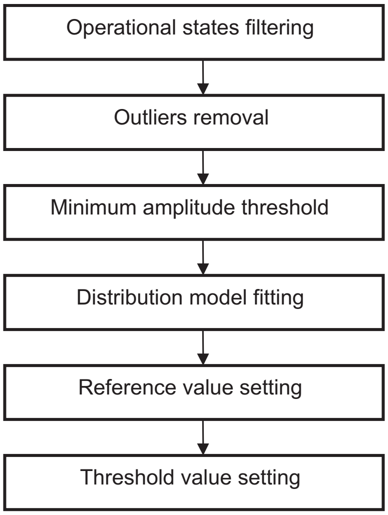
Operational states filtering
Outliers removal
Minimum amplitude threshold
Distribution model fitting
Reference value setting
Threshold value setting

Fig. 11. Flowchart of the automated threshold setting method.

acquisition caused either by software or hardware errors or external electrical or mechanical disturbances.

Regardless of the source, outliers are undesirable components, which ought to be filtered out a priori to modeling. The simplest technique is to remove outliers according to amplitudes, but this method is not feasible, since a different set of amplitudes is required for every estimator (trend). Another technique is to remove a certain percentage of the data, but the major pitfall of this approach is that in case of a large number of outliers, only a part of them will be removed. The technique suggested by the authors is to remove outliers according to a percentile value.[4] The industrial practice shows that (with an acceptable degree of imperfection), a single percentile value can be accepted for a particular machine.

In case of process values, percentiles should be set symmetrically, on both lower and upper border, whereas in case of a vibration data measures, the upper percentile ought to be much less than the lower percentile, e.g. of the order of 1:5. Actually, if the system can handle the computational burden, upper and lower percentiles might be adopted according to the goodness-of-the-fit. Outliers' removal becomes even more critical, if operational states are not defined, since a larger amount of data is not related to the process. Another benefit of percentiles application is that in case of a normally-distributed data, the disturbance introduced by their removal is practically insignificant.

5.4.3. Minimum amplitude threshold

For positive-amplitude data, like a vibration data measures level, it is often of an extra benefit to filter the data by a minimum threshold, which is accepted as a noise level. In this matter, a zero-value samples are also filtered out and do not require extra handling. It is especially suitable for low amplitude signals, where models built from highly noise-contaminated data are not permissible. If a minimum amplitude threshold is applied, a natural consequence is to apply a predefined minimum reference threshold level as well, i.e. reference threshold level applied if the level calculated from the data is lower.

5.5. Distribution model fitting

As already mentioned, the application of a model increases the chances to recover the true process behavior by smoothing potential randomness. In fact, the application of a model acts as a low-pass filter to the data – or (from a mathematical point of view) as an integrating, i.e. averaging tool. However, the application of a model requires an extra data preparation in addition to the ones described in Section 5.3. Consecutive steps are listed below.

5.5.1. Minimum number of data points

For a sound statistical model fitting, the first requirement is a sufficient number of data points. This is due to the fact, that statistics equations are often complex and require a relatively large data sets in order to converge plausibly. Moreover, statistical fits are calculated step-by-step, i.e. one coefficient after another, often engaging logarithmic operations, results of which might be truncated and lost if insufficient number of data points is applied.

5.5.2. Distribution function selection

A variety of statistical models are commonly available, e.g. Gaussian, gamma, beta, log-likehood, general extreme value, Weibull, Poisson, students, etc. Authors' research on different distribution functions has shown that for threshold limit setting the GEV (generalized extreme value) is generally the most suitable. The comparison of several distributions on several data sets will be presented in the next chapter. The GEV distribution acts similar to the Gaussian distribution, but it is much more flexible in a shape fitting. Nevertheless, it tries to keep the (usually desired) symmetry of the distribution.

As an alternative, the Weibull distribution can also accept asymmetric shapes (e.g. with exponential inclination) in contrast to the normal distribution. Such feature is of utmost importance for non-normally distributed data. The comparison in the next chapter will present that also for this type of data, the GEV is better than the Weibull distribution.

5.6. Setting the reference and threshold values

5.6.1. Reference value setting

The final step of the procedure converts the probability distribution into the threshold value. Often it is necessary to generate two levels: for alert and alarm limits. From theoretical point of view, it would be logical to set these limits to particular values of cumulative distribution functions (CDFs), for example 98% and 100%, respectively. The necessary equations for the calculation of CDF are usually available along with equations for the calculation of PDF. Our practice showed that such an approach still generates too many false alarms, especially for the vibration data measures. To avoid this practical problem, we found that it is beneficial to divide the threshold setting step into two steps.

The first step is to calculate a "reference value". It is a high percentile value of the fitted cumulative probability density. Typical values are in the range of 96–98%.

5.6.2. Threshold value setting

The last step is calculation of the final threshold value. It is defined as a given distance from the reference value. From various approaches possible, the authors have adopted the method recommended by [6], where the alert and alarm values are defined in dB spans.

Final remark concerning threshold settings is that for a process data (e.g. Fig. 7), it is suitable to set threshold levels symmetrically, as linear functions of the variance of a pre-processed data, but measured from reference percentile levels calculated from a model. For one-sided data (Figs. 5, 6 and 8), it is appropriate to set a one-sided threshold levels as a decibel (i.e. logarithmic) function of predefined increments (e.g. 3 and 6 dB for warning and alarm,

[4] Different software environments might handle percentiles in a slightly different way. It is always rational to verify the available commands.

respectively), measured from a reference percentile value calculated from a model.[5]

6. Case study – optimized fits on the real datasets

In this section the proposed algorithm was applied to four data sets, described previously in the Section 4. The first important problem was to estimate the accuracy of the fit. Next, several probability distributions were tested for the selected datasets.

6.1. Distribution function fit assessment

Once a model is applied, its degree of the fit to the data can be measured by so-called statistical ‘‘goodness-of-the-fit’’ tests. Clearly, these tests refer to assessment concerning a particular distribution; therefore, it is often more suitable to propose a general assessment of a model, which would point out the best fit from all of selected. In order to do that let us introduce a divergence measure between two distributions which is based on Hellinger distance [27]. The Hellinger-like distance measures a divergence between the empirical distribution P and a model distribution Q, and is given by following formula:

$$Hd = \sum_{t=1}^{t_{max}} (P(x_t) - Q(x_t))^2, \tag{8}$$

where $P(x_t)$ and $Q(x_t)$ denote empirical and fitted distributions respectively, and x_t denote signal samples. The Hellinger-like distance is a proper tool for comparison fitting goodness of various probability distributions for the same data. However, it is inconvenient for comparisons between various data sets, in particular for values of independent variables for which the widths of bins are significantly different. Therefore, in order to obtain a universal measure for various data sets, the percentage Hellinger-like distance (PHd) is introduced in following formula:

$$PHd = \frac{\sum_{t=1}^{t_{max}} (P(x_t) - Q(x_t))^2}{\sum_{t=1}^{t_{max}} P(x_t)^2}. \tag{9}$$

Both measures are to be applied in Section 6.2.

6.2. Comparison of selected probability distributions

In Section 4, examples of real data were introduced along with corresponding Gaussian distribution fits. It was shown that normal distribution fits well, if and only if the data is distributed close to normal, which in reality is not as common as one would expect. In Sections 3 and 5, the authors have laid down a set of rules to apply in data modeling. Section 6.2 presents resulting fits (for the same data as in Section 4). Values of Hd and PHd for preprocessed data sets for variables A, B, C and D are shown in Table 1.

Table 1
Accuracy of various distribution types fitting for four data sets.

Data set	Type of distribution	Hd (−)	PHd (%)
Set *A*	Extreme value	0.0036	15.13
	Gaussian	0.0015	6.30
	GEV	0.00085	3.57
	Inverse Gaussian	0.0009	3.78
	Weibull	0.0016	6.72
Set *B*	Extreme value	9.222	25.36
	Gaussian	6.900	18.98
	GEV	4.443	14.21
	Inverse Gaussian	5.150	14.21
	Weibull	6.700	18.43
Set *C*	Extreme value	0.0798	608.00
	Gaussian	0.0027	23.08
	GEV	0.0007	6.41
	Inverse Gaussian	0.0879	751.00
	Weibull	0.0029	24.79
Set *D*	Extreme value	11.8	10.51
	Gaussian	19.35	17.45
	GEV	11.25	10.15
	Inverse Gaussian	39.56	35.70
	Weibull	18.15	16.37

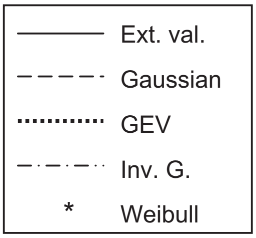

Fig. 12. Line styles legend for applied distribution functions.

According to the obtained results certain conclusions might be drawn:

- Virtually, Gaussian and Weibull distribution for sets A and B perform alike.
- The GEV distribution was found the best for all investigated cases.
- Other distributions might show accurate results only for a particular case (e.g. extreme value for set D).

Figs. 13–16 present four datasets after the processing according to the proposed method. Investigated probability distributions were also plotted. The algorithm-selected data is marked in black, and the rejected data is marked in grey. The legend illustrating applied distributions is shown in Fig. 12.

It is important to add that all the given data has been preprocessed according to states definitions. However, it is clear that this operation did not guarantee a normally distributed data. Only after implementation of described rules, the models enabled suitable threshold levels calculations.

In case of Figs. 14 and 16, the application of PDF does not represent the contents of the data sets in a suitable way, but it enabled a significant improvement comparing

[5] The article focuses on modeling of probability distribution functions. However, at this point is worth mentioning that the reference level is conveniently calculated from a cumulative distribution function.

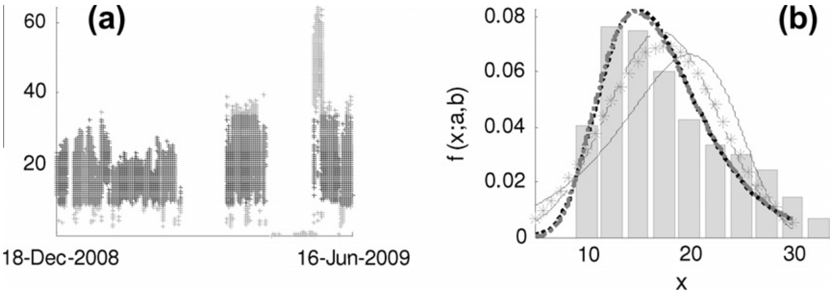

Fig. 13. Peak-to-peak trend value from a gas compressor (a) selected data trend, (b) selected data (normalized) histogram with investigated PDF fits.

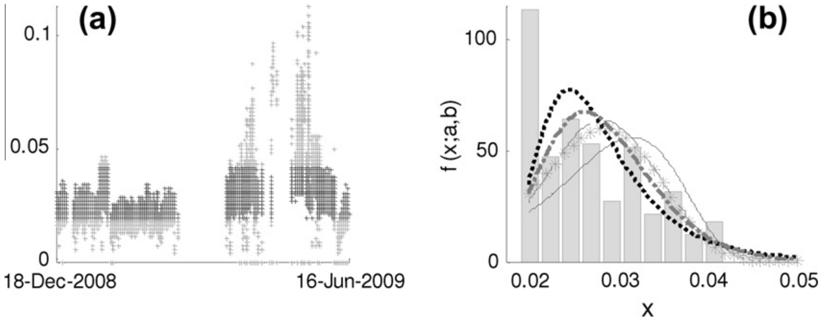

Fig. 14. Outer ring of a rolling element bearing from a gas compressor (a) selected data trend, and (b) selected data (normalized) histogram with investigated PDF fits.

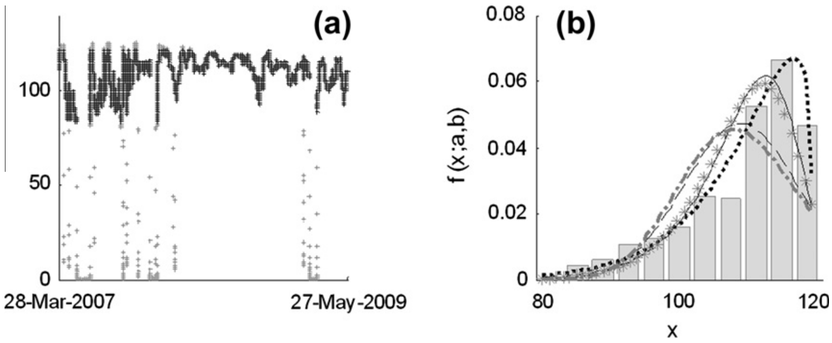

Fig. 15. Temperature from a gas compressor (a) selected data trend, (b) selected data (normalized) histogram with investigated PDF fits.

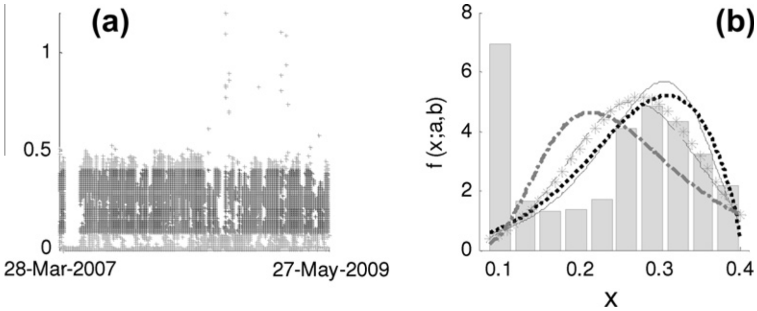

Fig. 16. Rolling element bearing from a wind turbine (a) selected data trend, (b) selected data (normalized) histogram with investigated PDF fits.

to raw data models. The mode values were not covered due to machines' performance, which was characterized by occasional vibration increase. Therefore, considering the dynamics of rotating machinery, this heavy left tail component should not be taken into account as primary reference level.

After conducted research, the authors came to following limitations of presented method:

- In some cases (less than 0.5%), proposed data preprocessing methods did not filtered the data sufficiently, and the data could not be modeled by proposed probability distribution functions at satisfactory level; such cases are to be handled separately.
- Although trend values could be grouped, and extra dependencies could be taken into account (e.g. for the same machine or for the same channel), the paper handles all trend data independently; an extra research concerning this matter is currently conducted.
- The method assumes that it operates either on healthy-machine-data or data with occasional failure signatures; it is not designated for explicit faulty-machine-data.

7. Conclusions

In the paper, it was shown that application of statistical models improves the quality of a raw data by diminishing the influence of potential random quantities, plus it enables the implementation of model-based algorithm for threshold settings. The paper presents numerous tips for real data modeling concerning data preprocessing, model selection, and threshold settings. Furthermore, the authors have emphasized the importance of diverse approaches towards data of different characteristics, namely a symmetrical data, and one-sided, positive data. Finally, presented illustrations show the true benefits of application of presented method. Although accompanied by a number of statistical formulas, the paper might be beneficial for practical engineers and researchers as well, since it gives a number of solutions to data management problems encountered in industry.

Apart from threshold levels applied to trend series illustrated in the paper, condition monitoring systems might apply limits to spectral and multidimensional representations of signals. In further works, the authors would like to extend the concept of automatic threshold settings to advanced multidimensional analysis like presented in [28–30], which still require a visual interpretation for diagnostic reasoning. Apart from presented approach, this work will require additional pattern recognition and possibly image processing techniques.

Acknowledgement

The authors would like to acknowledge the financial support of the Polish Ministry of Science and Higher Education under the research Grant No. N504670540.

References

[1] T. Gellermann, Requirements for condition monitoring systems for wind turbines, AZT Expertentage, 10–11.11.2003, Allianz, 2003.
[2] Germanischer Lloyd, Richtlinien für die Zertifizierung von Condition Monitoring Systemen für Windenergieanlagen, Vorschriften und Richtlinien, Selbstverlag Germanischer Lloyd, 2007.
[3] VDI 3834 – Messung und Beurteilung der mechanischen Schwingungen von Windenergieanlagen und deren Komponenten (Measurement and evaluation of the mechanical vibration of wind energy turbines and their components), Verein Deutscher Ingenieure, March 2009.
[4] T. Barszcz, Selection of diagnostic algorithms for wind turbines, in: Proceedings of 4th International Congress on Technical Diagnostics, Olsztyn, Poland, 2008.
[5] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 2000, ISBN: 0-471-05669-3.
[6] C. Cempel, Simple condition forecasting techniques in vibroacoustical diagnostics, Mechanical Systems and Signal Processing 1/1 (1987).
[7] C. Cempel, Limit value in practice of vibration diagnosis, Mechanical Systems and Signal Processing 4/6 (1990).
[8] A. Heng, S. Zhang, A.C.C. Tan, J. Mathew, Rotating machinery prognostics: state of the art, challenges and opportunities, Mechanical Systems and Signal Processing 23/3 (2009).
[9] P. Naga, S. Rao, A. Naikan, An algorithm for simultaneous optimization of parameters of condition-based preventive maintenance, Structural Health Monitoring 8 (2009).
[10] R. Brooks, R. Thorpe, J. Wilson, A new method for defining and managing process alarms and for correcting process operation when an alarm occurs, Journal of Hazardous Materials 115 (2004).
[11] M. Bransby, J. Jenkinson, The management of alarm systems, HSE contract research report 166/1998, HSE Books, 1998, ISBN 0-7176-1515 4.
[12] W. Bartelmus, R. Zimroz, A new feature for monitoring the condition of gearboxes in non-stationary operation conditions, Mechanical Systems and Signal Processing 23 (2009) 1528–1534.
[13] W. Bartelmus, R. Zimroz, Vibration condition monitoring of planetary gearbox under varying external load, Mechanical Systems and Signal Processing 23 (2009) 246–257.
[14] Zimroz et al., Measurement of instantaneous shaft speed by advanced vibration signal processing – application to wind turbine gearbox, Metrology and Measurement Systems 18/4 (2011) 701–712.
[15] K. Spackman, Signal detection theory: valuable tools for evaluating inductive learning, in: Proceedings of the 6th International Workshop on Machine Learning, San Mateo, 1989.
[16] J. Huang, C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Transactions on Knowledge and Data Engineering 17/3 (2005).
[17] R. Pincus, V.K. Rohatki, Statistical Inference, John Wiley & Sons, 1984 (Biometrical Journal 27/4 (1985)).
[18] P. Ramirez, J.A. Carta, Influence of the data sampling interval in the estimation of the parameters of the weibull wind speed probability density distribution: a case study, Energy Conversion and Management 46 (2005).
[19] Statistics ToolboxTM User's Guide, The MathWorks Inc., Version 4.0, 2002, pp. 2–23.
[20] P. Rousseeuw, A. Leroy, Robust Regression and Outlier Detection. John Wiley & Sons, 2003. ISBN: 978-0-471-48855-0.
[21] D.M. Hawkins, Identification of Outliers, Chapman and Hall, 1980, ISBN 0-412-21900-X.
[22] S. Ross, Peirce's criterion for the elimination of suspect experimental data, Journal of Engineering Technology 20/2 (2003).
[23] B. Iglewicz, D. Hoaglin, How to Detect and Handle Outliers, ASQC Quality Press, 1993, ISBN 0-873-89247 X.
[24] B. Rosner, Percentage points for a generalized ESD many-outlier procedure, Technometrics 25/2 (1983).
[25] J. Taylor, An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements, second ed., University Science Books, 1997, ISBN 0-935702-75-X.
[26] R. Dean, W. Dixon, Simplified statistics for small numbers of observations. Analytical Chemistry 23/4 (1951).
[27] N. Sazuka, On the gap between an empirical distribution and an exponential distribution of waiting times for price changes in a financial market, Physica A: Statistical Mechanics and its Applications 376 (2007).

[28] J. Urbanek, J. Antoni, T. Barszcz, Detection of signal component modulations using modulation intensity distribution, Mechanical Systems and Signal Processing 28 (2012) 399–413.

[29] A. Jablonski, T. Barszcz, Instantaneous circular pitch cyclic power (ICPCP) – a tool for diagnosis of planetary gearboxes, Key Engineering Materials 518 (2012) 168–173.

[30] Y. Liu, Bispectrum analysis for feature extraction of pitting fault in wind turbine gearbox, in: Proceedings of the 2010 IEEE International Conference on Mechatronics and Automation August 4–7, Xi'an, China, 2010.