IOP Conf. Series: Materials Science and Engineering **289** (2018) 012032 doi:10.1088/1757-899X/289/1/012032

Improvement of statistical methods for detecting anomalies in climate and environmental monitoring systems

A G Yakunin[1], H M Hussein[2]
[1]Head of chair, Altai State Technical University, Barnaul, Russia
[2]Assistant professor, Minia University, Minia, Egypt
E-mail: yakunin@agtu.secna.ru

Abstract. The article shows how the known statistical methods, which are widely used in solving financial problems and a number of other fields of science and technology, can be effectively applied after minor modification for solving such problems in climate and environment monitoring systems, as the detection of anomalies in the form of abrupt changes in signal levels, the occurrence of positive and negative outliers and the violation of the cycle form in periodic processes.

1. Introduction

Recently, continuous on-line monitoring with the ability to automatically detect emergency situations is gaining relevance in solving many technical and environmental monitoring applications. However, existing solutions for identifying such situations are quite limited possibilities [1–5].

Furthermore, with low-cost devices for primary data collection, the volume of measurements has increased dramatically that has led to the problem of its timely stream processing, transmission, storing, and its immediate access [6–10].

In this context, it becomes urgent to develop software-hardware and algorithmic solutions to effectively detect anomalies, i.e., violations of laws in controlled processes based on their features and minimize the amount of transmitted and stored information without losing its pragmatic values. As was shown in [11–13], among all types of anomalies in monitoring systems, the most common and unambiguously interpreted are the fluctuations, outliers and cyclical violations of the recorded signals in monitoring systems, and statistical methods are the most suitable for their identification. And in the development of data compression algorithms can be used such a signal property as smooth character of their changes.

This paper deals only with issues devoted to solving the problem of identifying the most characteristic anomalies. It offers various options for improvements of well-known statistical methods of processing the data and their application for solving such problems of climate and environmental monitoring systems, as the detection of violations of the regularities of the flow of controlled processes.

2. Model, methodology and equipment of the experiment

The proposed methodology is a computational experiment conducted on model signal capable to reproduce the previously listed anomalies types. The source data for determining the parameters of the model signal were obtained by using measured data of technical and meteorological monitoring system in campus of Altai State Technical University. Hardware and software used in the system have repeatedly described in [14–19]. The same system was used for full-scale testing and effectiveness evaluation of the developed algorithmic solutions.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

Since a large amount of information in the system was the temperature data, special attention was paid to data processing from temperature sensors.

To investigate the algorithms for detecting anomalies and compressing data using a computational experiment, the following model has been proposed to describe the measured signals from sensors:

$$x(t) = x_s(t)+x_p(t) \cdot k_{am}(t)+x_a(t)+x_n(t), \quad (1)$$

wherein $x_s(t)$ – non-periodic quasi deterministic component of a useful informative signal, which is not generally a random process and provides information about changing the value of the controlled parameter over time; $x_p(t)$ – a periodic component of the information signal; $k_{am}(t)$ and $x_a(t)$ – respectively, multiplicative and additive components of the abnormal and $x_n(t)$ – the noise component.

To describe the informative signal, a model function was used in the form of sine wave with an angular frequency ω_o and amplitude A_o:

$$x_s(t) = A_o \cdot Sin(\omega_o \cdot t). \quad (2)$$

The frequency of this sine wave $f_o = \omega_o / 2\pi$ was selected about 50 ... 5000 times lower than the sampling frequency of the signal $x(t)$. A higher frequency was used when there was a periodic component in the signal. Signal amplitude A_o determines the maximum possible rate of rise or fall of the signal, i.e., limit for its trend.

The component $x_p(t)$ was used for modeling the periodic processes. It was implemented using two types of modeling functions: sinusoidal and triangular shapes with the same frequency f_p, and amplitude A_p:

$$x_p(t) = A_p \cdot Sin(2\pi f_p \cdot t); \quad (3)$$

$$x_p(t) = F(t) * A_p \cdot (t/2f_p - 2i), \text{ if } i/f_p < t <= (i+0.5)/f_p; \quad (4a)$$

$$x_p(t) = F(t) * A_p \cdot (2i - t/2f_p), \text{ if } (i+0.5)/f_p < t <= (i+1)/f_p, \quad (4b)$$

where i – nonnegative integer that specifies the number of triangular "teeth", the sign "*" means convolution function and smoothing sharp peaks – conversion is described by Gaussian function with a time constant of $\tau = 0.1 \cdot f_p$:

$$F(t) = 1/(\tau \sqrt{2\pi}) \cdot exp(-t^2/2 \tau^2) \quad (5)$$

Cycles frequency f_p was chosen 20–100 times higher than the frequency F_o and also cycle time had to at least 20–100 counts.

The function of a rectangular window with an amplitude lower unity was used to describe the multiplicative abnormal component. The beginning and end of the window coincide with the beginning and end of the ***i-th*** period:

$$k_{am}(t) = k_{mo} \cdot rect (f_p \cdot t - i + 0.5), \quad (6)$$

where $rect(t) = 0$ for $|t| > 0.5$, 0.5 when $|t| = 0.5$ and 0 otherwise.

The additive type component of abnormal depended on the type of anomaly. In the simplest case for single level difference by the amount A_a occurring at time t_o, it was represented as a Heaviside function:

$$x_a(t) = A_a \cdot h (t - t_o), \quad (7)$$

where $h(t - t_o) = 1$ for $t > t_o$, 0.5 at $t = t_o$ and to 0 otherwise.

For a more accurate description of the real signals level changes, an approximation of the following form was used

$$x_a(t) = F(t) * \{A_a \cdot [(t - t_o\, sign(A_a)) / \tau_f + 0.5] \cdot rect (t - t_o / \tau_f)\}, \quad (8)$$

where τ_f – the length of the front level changes, and the function $F(t)$ described by equation (5), wherein the constant τ was selected to be from 0.1 to 0.3 of the duration of τ_f. The duration of the front level changes themselves varied within 3 ... 20 times the sampling rate of the signal Δt.

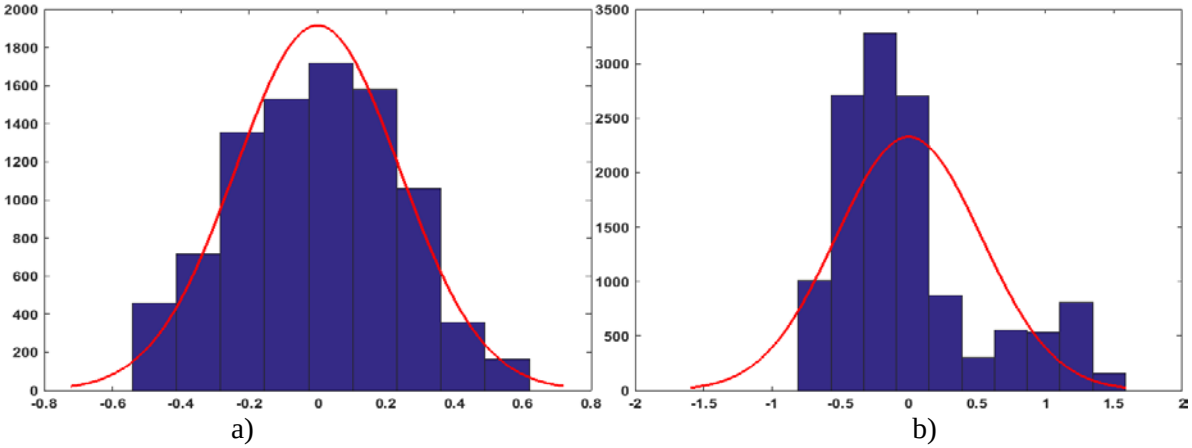
In order to accumulate statistics on error detection of level changes, they interfered with the information signal model in the form of a series in which the interval between the level changes varied according to the uniform law between the minimum and maximum values, with constant alternation of positive and negative level changes.

The minimum repetition time for level changes lay in the range of 0.5 to three periods of the information signal x_s, but not less than 30 ... 100 durations level changes fronts. The maximum repetition time exceeded the minimum value by 2– 3 times so that the total duration of a series of observations took a reasonable time.

For outlier modeling, it is either used the Gaussian Function $F(t)$ described by expression (5) multiplied by the required amplitude of level deference A_a, or the outlier can be formed of two bipolar level changes with same amplitude as described in the expression (7) or (8) and the interval between them was chosen to be within 10 -50 times the signal sampling rate.

To select an adequate model to describe the noise component of the signal $x_n(t)$ describing the fluctuation of the information signal due to the influence of many difficulties of the destabilizing factors, series of field experiments was carried out to determine its statistical properties. Since, according to the model (1) it was assumed that the dynamic impact of such factors is much higher than dynamics of informative component and is comparable with the dynamics of changes for abnormal components, to extract the informative component, the signal was subjected to low-frequency filter by double-smoothing it with a rectangular window of 50–100 samples width, and the result was subtracted from the original signal. A smaller width of the smoothing window was used to analyze periodic signals when the length of the cycle contained about 500 samples. If the number of samples was smaller, a signal averaged over 10–20 periods was subtracted from each period of the original signal (2–10 before and the same after the current period). Thus, regions that contained anomalies and that did not contain were analyzed separately.

As a result of the conducted studies, it was found that in the absence of anomalies, the properties of the stochastic signal component extracted in the manner described above are close to the properties of normal white noise, since the distribution law of the readings obtained is close to normal, and the autocorrelation function has a well-defined peak. As an example, Figure 1 shows the distribution laws of $x_n(t)$ samples for the temperature graphs and graphs of pressure changes.



a) b)

Figure 1. Distribution histogram of temperature random fluctuations in the absence (*a*) and existence (*b*) of anomaly. The width of the smoothing window – 50 samples, a series of observations on 07–10 February 2014.

As seen from Figure 1, for an anomaly-free portion, the distribution law is close to normal, whereas, in the presence of anomalies on the side lobes of the distribution histogram, additional ups are appearing which is quite natural, since the anomalies are a non-stationary and nonergodic process.

The described model was used to carry out computational experiments to investigate proposed and modified known processing algorithms, which consisted in analyzing the dependencies of identification errors of various types of anomalies for their parameters, information signal parameters, signal-to-noise ratio and parameters included in the detection algorithm. The ultimate goal of these studies was the parameters selection for algorithms and estimating their possible effective application. Only at the final stage of the study, computational experiment was replaced by field experiment in order to confirm the results obtained in the computational experiment, and their comparison with reference methods of identification. Let us first discuss in detail algorithms for identifying violations of regularity of controlled processes.

3. Algorithm for detection downs and ups of signal level

The essence of the proposed algorithm (Table 1) is based on comparing of average speeds changes in adjacent with equal time intervals and containing $\boldsymbol{n}$ samples into which the results of measurements of *x (t)* informative signal was divided. the average value of controlled variable is computed for each of these intervals. Then, with these values, the averaged derivative of *x (t)* is calculated using the two-point scheme, that is actually average speed of change of controlled parameter on time. The threshold tolerance value $\boldsymbol{\delta}$ of the speed deviation is calculated as a modified standard deviation of speed at individual intervals. Depending on the ratio of threshold tolerance value and the current change rate of the signal, the analyzed time zone can be classified into a normal, critical zone or a zone of anomalies.

Table 1. Algorithm for identifying anomalies in the form of signal level changes.

Calculating the average value of the measured data $\boldsymbol{x(t)}$ at the k-interval that includes $\boldsymbol{n}$ discrete samples::	$\boldsymbol{x_k} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=(k-1)*n+1}}^{\boldsymbol{k*n}} \boldsymbol{x(t_i)}$; (9)
Calculating the difference $\boldsymbol{v}$ of the average values between adjacent intervals (or between the maximum and minimum of the candles):	$\boldsymbol{v_k} = (\boldsymbol{x_k} - \boldsymbol{x_{k-1}})/\boldsymbol{n}$ **or** $\boldsymbol{v_k} = (\boldsymbol{high_k} - \boldsymbol{Low_k})$ (10)
Calculating the average $\overline{\boldsymbol{v}}$ for the difference $\boldsymbol{v}$ in $\boldsymbol{n}$ intervals:	$\overline{\boldsymbol{v_k}} = \frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=k-N+1}}^{\boldsymbol{k}} \lvert\boldsymbol{v_i}\rvert$; (11)
Calculating the deviation $\boldsymbol{\Delta v}$ of speed changes:	$\boldsymbol{\Delta v_k} = \boldsymbol{v_k} - \overline{\boldsymbol{v_k}}\ \boldsymbol{.sign(v_k)}$ (12)
Calculating the threshold of tolerance $\boldsymbol{\delta}$:	$\boldsymbol{\delta} = \frac{\boldsymbol{1}}{\boldsymbol{N}}\sqrt{\sum_{\boldsymbol{i=k-N+1}}^{\boldsymbol{k}} \boldsymbol{\Delta v_k^2}}$ (13)
Making a decision on the admissibility of $\boldsymbol{\Delta v}$ values:	

$$\forall \boldsymbol{\Delta v_k}: \begin{cases} \lvert\boldsymbol{\Delta v_k}\rvert \le \boldsymbol{\delta} & : \text{Normal zone,} \\ \boldsymbol{\delta} < \lvert\boldsymbol{\Delta v_k}\rvert < \boldsymbol{k_\varepsilon \delta} & : \text{Critical zone ,} \\ \lvert\boldsymbol{\Delta v_k}\rvert \ge \boldsymbol{k_\varepsilon \delta} & : \text{Zone of anomalies.} \end{cases} \quad (14)$$

As an example, Figure 2*a* shows the result of applying the proposed algorithm to real data coming from digital thermal sensor of an information system. To assess the effect of noise on the value of the probability of errors (as type I and type II errors) of identifying level changes, additional studies have been conducted, the results of which are shown in Figure 2*b*. And Figure 2*c* shows comparison of noise effect on error of level changes detection for Bayesian method, which is often used for similar purposes. As seen from the Figure 2*c*, the Bayesian method is more sensitive to noise than proposed.

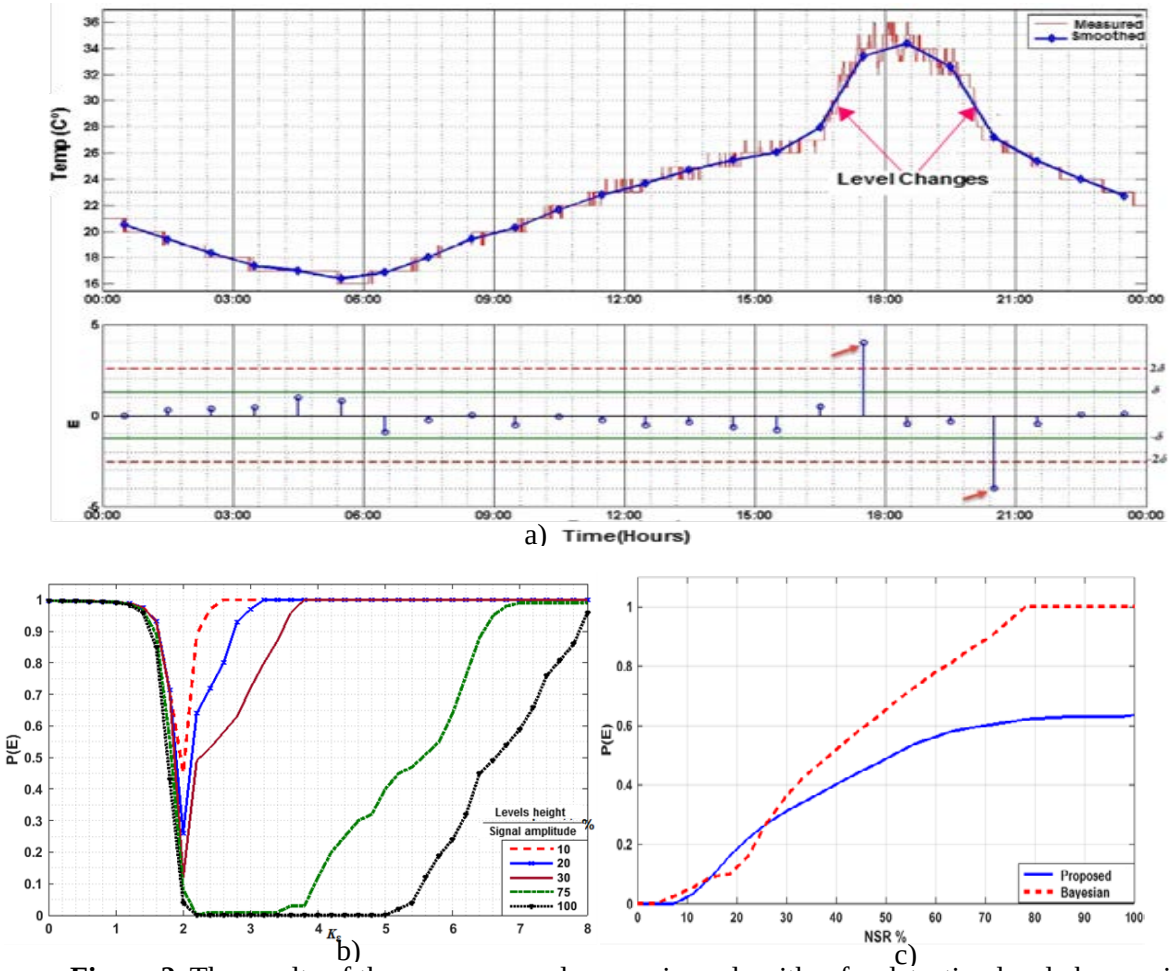


Figure 2. The results of the average speed comparison algorithm for detecting level changes in temperature (*a*), dependence of the error probability for level change identification on kε for different level change amplitudes (*b*) and the dependence of the error probability on the signal to noise ratio for the proposed and Bayesian methods (*c*).

4. Algorithm for detection signal outliers

To detect outliers, a method has been developed based on the analysis of clusters in the trend of the time series. The method is able to identify two types of outliers: type A when observed anomalous trend lies above the expected measured data boundaries and Type B when the emission falls below the expected limits. In contrast to the level changes detection method, in this case, the changes of average speed for adjacent intervals is analyzed by the three-point scheme:

$$a_k = (x_k + x_{k-2} - x_{k-1}), \tag{15}$$

The physical analogue of a_k is the change in the magnitude of the function's trend, that is, the acceleration. In contrast to (11), in expression (15) division into the number of samples N in the average interval is omitted, which reduces the computational complexity, but does not affect the performance of the method. All the other equations of average speed comparison method (12) ... (14) valid with the only difference that the symbol of speed $\boldsymbol{v}$ should be replaced with the acceleration symbol $\boldsymbol{a}$:

$$\bar{a}_k = \frac{1}{N}\sum_{i=k-N+1}^{k} |a_i|. \tag{16}$$

$$\Delta a_k = a_k - \bar{a}_k sign(a_k), \tag{17}$$

$$\Delta a = \delta = \frac{1}{N}\sqrt{\sum_{i=k-N+1}^{k} \Delta a_k{}^2}. \tag{18}$$

$$\forall \Delta a_k : \begin{cases} |\Delta a_k| \leq \delta : Normal\ zone, \\ \delta < |\Delta a_k| < K_\varepsilon \delta : Critical\ zone; \\ |\Delta a_k| \geq K_\varepsilon \delta : Zone\ of\ outliers. \end{cases} \tag{19}$$

According to the mathematical analysis rules, if the second derivative of a function is positive, and the first equals zero, then there is a minimum, otherwise (a <0 and v = 0) is the maximum of the function. However, the control of the condition v = 0, reducible to the given case, to checking divergence of the signs of average speeds $\boldsymbol{v_k}$ and $\boldsymbol{v_{k_-1}}$ on the last two intervals is not usually required, since when the last two inequalities in system (19) are satisfied, this condition is practically always satisfied. An example of operation of the proposed algorithm is shown in Figure 3.

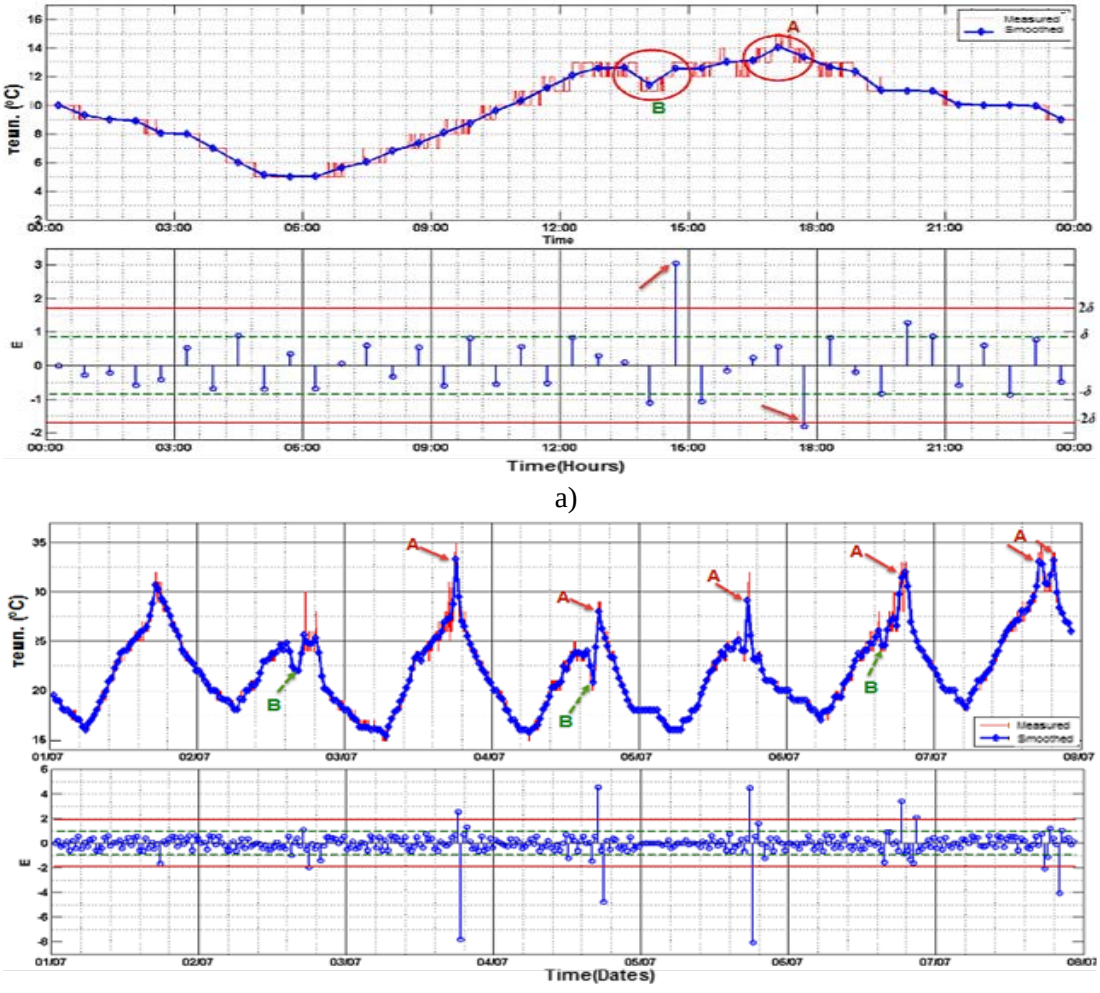


Figure 3. Detection of outliers' type B and A on daily fluctuations of outdoor temperature on 20 June 2014 (*a*) and during a week from 1 to 7 July 2014 (*b*)

In order to objectively evaluate the effectiveness of the proposed method, a computation experiment was conducted using the developed model of the information signal. The dependence of the total error of outliers' detection (type I and II), depending on the signal to noise ratio for various values of the outliers' amplitudes to the signal amplitude ratio, the ratio of the outlier duration to the period of harmonic oscillations, and the various time intervals of the observation was investigated. As an exam-

ple, Figure 4 shows such dependencies for the ratio of amplitudes is equal to 0.2 and the ratio of periods is equal to 0.1. The period of harmonic oscillations contained 250 samples.

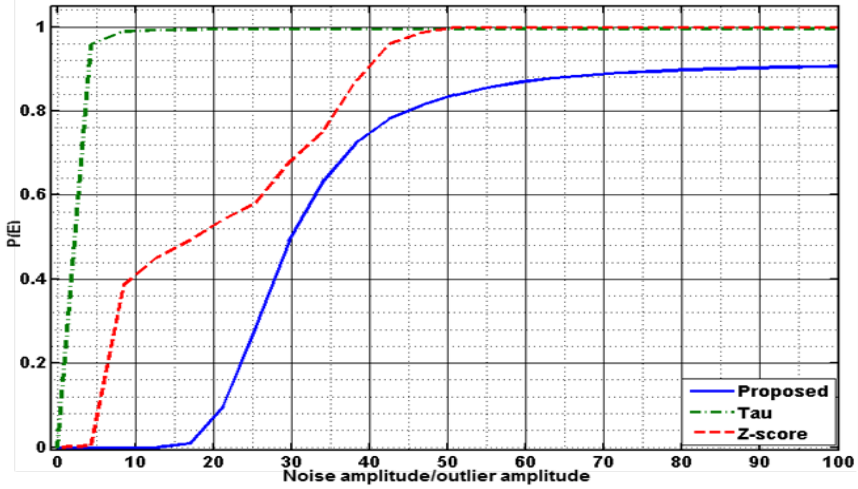


Figure 4. Dependence of outliers' detection error on signal to noise ratio (in %) for the proposed method based on trend comparison, Z-score method and τ – Thompson method. A_O / A_a = 30%.

For comparison, the same Figure 4 shows the dependencies for the methods Z-score [20,21] and τ – Thompson [22], which are also widely used for outlier detection. However, these methods did not work efficiently without certain modifications. For example, in Z-score method after normalization of original sequence according to the formula [21]

$$y_i = \frac{0.6745[x(t_i)-x_m]}{y_m}, \tag{20}$$

where $x_m = median[x(t_i)]$– median value of the analyzed sequence, and $y_m = median[|x(t_i)- x_m|]$ – normalizing coefficient, the obtained values were compared not with a fixed value but with a floating threshold given by the expression

$$y_o = median[||x(t_i)- x_m|- y_m|] \tag{21}$$

As seen in Figure 4, the proposed algorithm has clear advantages, particularly at high noise levels.

5. Algorithm for distortions detection of periodical signals

Many controlled processes in the monitoring systems are characterized cyclic behavior. It is therefore very important to detect in a timely and reliable way the deviations of the process's real behavior from its pattern. To detect these cycling violations, a modification of the method based on the pattern forms was proposed, described by the following algorithm [12].

Step 1. Normalization of the measured data on the cycle period by Z-scores:

$$y(t) = \frac{x(t)-\overline{x(t)}}{\sigma}; \tag{22}$$

Step 2. Finding a trend for an aperiodic signal component $x_s(t)$:

$$x_{s_i}^{k-1} = \begin{cases} \frac{\frac{m}{2}-i}{m}\bar{x}_{k-2} + \frac{\frac{m}{2}+i}{m}\bar{x}_{k-1} \; for \; i = 1 \ldots \frac{m}{2} \\ \frac{3m-2i}{2m}\bar{x}_{k-1} - \frac{\frac{m}{2}-i}{m}\bar{x}_k \; for \; i = \frac{m}{2}+1 \ldots m \end{cases} \tag{23}$$

Step 3. Recovery values purely periodic signal component:

$$x_{p_i}^{k-1} = x_i^{k-1} - x_{s_i}^{k-1} \tag{24}$$

Step 4. Evaluation of the signal amplitude in the previous cycle:D_{k-1}

$$D_{k-1} = max_{i=1\ldots m}\, x_{p_i}^{k-1} - min_{i=1\ldots m}\, x_{p_i}^{k-1} \quad (25)$$

Step 5: Detection of cyclicity violations if the condition $\boldsymbol{k > 3}$ is satisfied:

$$max_{i=1\ldots m}\left(\frac{x_{p_i}^{k-1}}{D_{k-1}} - P_i^{k-2}\right) > \delta\,, P_i^{k-1} = \alpha\frac{x_{p_i}^{k-1}}{D_{k-1}} + (1-\alpha)P_i^{k-2}.$$
$$\overline{D}_{k-1} = \alpha D_{k-1} + (1-\alpha)\overline{D}_{k-2} \quad (26)$$

where $\boldsymbol{\delta}$ is the detection criterion specifying the value of the maximum permissible relative deviation, $\boldsymbol{\alpha}$ is the exponential smoothing coefficient, approximately equals to $\frac{1}{n}$.

The algorithm includes data smoothing, normalizing them, forming a data period template with non-zero coefficients in the trend line, calculating the actual deviation of the actual data from the template, and comparing the deviation with the acceptable threshold. If the cycle period is unknown, before applying the algorithm it must first be found, for example, using the discrete Fourier transform. Unlike other similar solutions, during the formation of the template, a compensation for the trend of the monitored signal was applied, which allowed increasing the authenticity of the identification.

The results of practical application of the method for detecting violations in cyclic processes are shown in Figure 5. For calculations, the temperature values measured every 30 seconds were taken. The measured values were smoothed using the moving average method with an hour window. The normalized data were divided into one-day periods that were used to form the primary template and calculate the correction factor.

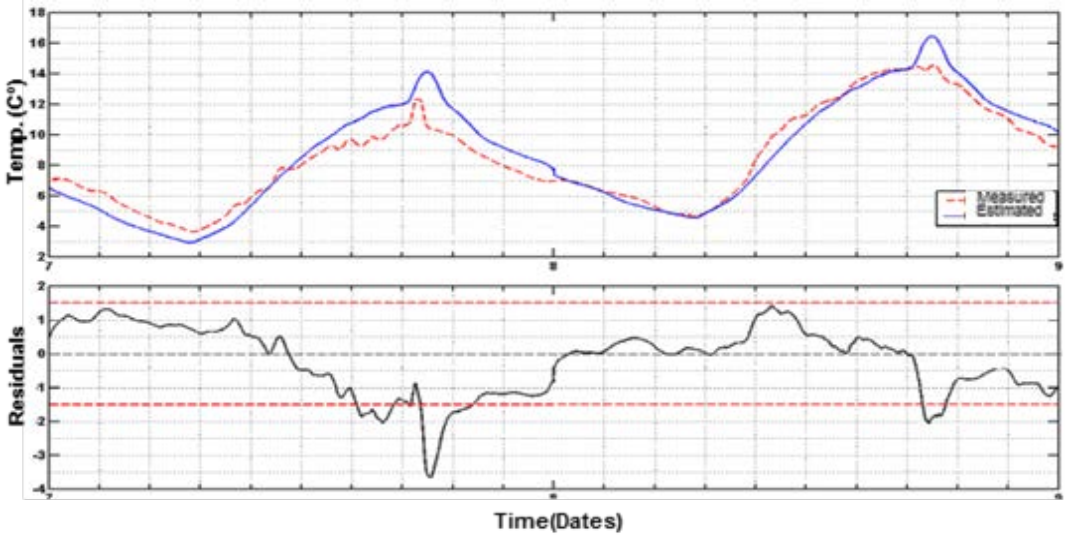


Figure 5- The original data, the corresponding patterns and the relative normalized differences between the original signal and the patterns.

6. Conclusions

In the course of the studies it was found that the use of the proposed modifications of known statistical methods for the detection of anomalies makes it possible, because of their low computational complexity, it is easy to implement methods using the simplest computing equipment in the vicinity of the sensor. As a result, in some cases it is no longer necessary to transfer the entire stream of recorded information, replacing it with the transmission of information only at intervals of the time on which the anomaly is occurring. At the same time, it is important to note that the reliability of the detection of anomalies in the proposed methods is as good as, and often better than the reference methods. All this makes it very promising to apply the proposed solutions in different systems not only to environmental monitoring or data collection systems but also to technical monitoring systems as well as to SCADA systems of various types.

Acknowledgement
The authors express deep gratitude to the management of Altai State Technical University. II Polzunov for financial support of this publication and OOO "Promavtomatika" for technical assistance in the preparation of article, as well as former students of the department computer science, computer engineering and information security department of the aforementioned University for their active participation in the development of hardware and software, which has allowed to realize the described researches.

References
[1] Ye M, Li X and Orlowska M E 2009 *Expert Syst. Appl.* **36** 7104–7113 doi: 10.1016/j.eswa.2008.08.030
[2] Huang Y, Hu L, Yang D and Liu H 2017 *IOP Conf. Ser. Earth Environ. Sci.* **81** 012208 doi: 10.1088/1755-1315/81/1/012208
[3] Branch J, Szymanski B, Giannella C et al 2006 *Proceedings - International Conference on Distributed Computing Systems* **2006** 1648838 doi: 10.1109/ICDCS.2006.49
[4] Ruggieri E 2013 *Int. J. Climatol.* **33** 520–528 doi: 10.1002/joc.3447
[5] Subramaniam S, Palpanas T, Papadopoulos D, Kalogeraki V and Gunopulos D 2006 *VLDB '06 Proc. 32nd Int. Conf. Very large data bases* 187–198
[6] Graefe G and Shapiro L D *Computer Science Symp. on Applied Compurting, Kansas City,Apr, 1991 - Proceedings of ACM/IEEE, 199.* 22–27
[7] Zhen C and Ren B 2009 *CiSE 2009* 5366670 doi: 10.1109/CISE.2009.5366670
[8] Ester M, Kriegel H P, Sander J and Xu X 1996 *Second International Conference on Knowledge Discovery and Data Mining* pp 226–31 URL: http://www.cse.ust.hk/~raywong/comp5331/References/ADensityBasedAlgorithmForDiscoveringClustersInLargeSpatialDatabasesWithNoise.pdf
[9] Hussein H M and Yakunin A G 2013 *Polzunovsky Vestn.* 65–68 (in Russian)
[10] Hussein H M, Suchkova L I and Yakunin M A 2012 *Polzunovsky Alm. AltSTU* 48–50 (in Russian)
[11] Hussein H and Yakunin A 2015 *Mater. XI Int. Res. Pract. Conf. «Modern Sci. potential-2015». Tech. Sci. - Sheff.* **38** 33–38 (in Russian)
[12] Hussein H S and Yakunin A G 2015 *J. Sib. Fed. Univ. Math. Phys.* **8** 157–164
[13] Hunt B 2007 *Int. J. Climatology* **27** 139–156 doi: 10.1002/joc.1379
[14] Hussein H M, Kuntz R V, Suchkova L I, Yakunin A G and Yurchenko A V 2013 *IV Scientific-practical conference "Information and measuring equipment and technology"* 160–169 (in Russian)
[15] Hussein H M, Kuntz R V, Suchkova L I and Yakunin A G 2013 *Izv. Altai State Univ.* **1** 210–4
[16] Suchkova L, Hussein H M, Yakunin M A, Yakunin A G and Yurchenko A V 2013 *Kontrol'. Diagnostika* **13** 95–101(in Russian)
[17] Yakunin A G 2017 *J. Phys. Conf. Ser.* **881** 012030 doi: 10.1088/1742-6596/881/1/012030
[18] Suchkova L I, Hussein H M, Yakunin M A and Yakunin A G 2013 *Polzunovsky Alm.* 75–82
[19] Yakunin A G and Hussein H M 2017 *J. Phys. Conf. Ser.* **881** 012028 doi: 10.1088/1742-6596/881/1/012028
[20] Shiffler R E 1988 *Am. Stat.* **42** 79–80 doi: 10.1080/00031305.1988.10475530
[21] Iglewicz B and Hoaglin D C 1993 *How to Detect and Handle Outliers* (United States:ASQC Quality Press)
[22] Thompson W R 1935 *Ann. Math. Stat.* **6** 214–219