УДК 621.3.08; 658.562; 620.179.1

OUTLIERS DETECTION IN AIR TEMPERATURE MEASUREMENTS

H. M. Hussein, A. G. Yakunin

Outliers are anomalous readings within the measured data set. Whatever the cause of it, which is numerous, it must be detected and eliminated for accurate assessment of the expected behavior. The current work developed novel methods to detect outliers in air temperature measurements in weather monitoring system. Moving average change rate method and candlestick graph. Both methods were applied to a random sample of air temperature measurements. They categorized the measured data into three zones: normal, suspected and outliers. Another three famous methods for outliers' detection were reviewed and compared to the proposed methods. The comparison showed that, the proposed methods detected the outlier boundaries simply and accurately.

Keywords: *Outliers; Moving average; Candlestick chart; modified Z-score; modified Thompson tau; modified boxplot.*

Introduction

An outlier is an observation point that deviate from other observations [1].

In [2] outlier has been considered as an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.

Outliers may be occurred due to variability in the measurement environment or it may indicate experimental error; the latter arise due mechanical faults, changes in system behavior, fraudulent behavior and human error. They are sometimes excluded from the data set.

Outlier detection aims to find patterns in the measured data that do not conform to expected behavior. It has extensive use in many applications. But are not rigid mathematical definitions for constituting outliers; determining whether or not an observation is an outlier is ultimately a subjective exercise. Many researches propose different techniques for outlier detection [3-5].

These techniques include Parametric Statistical Modeling [6-9], Neural Networks [10-12], Spectral Techniques [13], Nearest Neighbor Based Techniques [14,15], Bayesian Networks [16,17] and more.

This paper presents new methods for detecting outliers in weather monitoring, specifically in air temperature measurements.

These methods based on change rate for moving average and candlestick graph are proposed to improve the accuracy of outliers' detection for air temperature measurements. However, they may be generalized to cover different observations.

The following sections will describe in detail the proposed methods.

1. Change rate for moving average method.

The moving average[18] depends on dividing the measured data ***f(t)*** into symmetrical time slots "***s***". Then, the average in every time slot will be calculated according to the following equation:

$$y_k(\ddot{t}) = \text{mean}\ (f_i(t) : f_{i+s}(t)) \qquad (1)$$

Where: *k:1,2,……,n(the number of time slots),* $\ddot{t}$ *= mean(t_i : t_{i+s}),i = 1, s, 2s, ……….., N-s, N : the size of f(t) and s is the time slot width.*

After calculating the slots averages, the temperature change rates $\bar{y}(\ddot{t})$ will be calculated as the difference between each successive two points using the following equation:

$$\bar{y}_k(\ddot{t}) = (y_k(\ddot{t}) - y_{k-1}(\ddot{t}))/s \qquad (2)$$

Where: $\bar{y}_1(\ddot{t})=0$, *k=2:n.*

The average of the change rates ***d*** will be calculated using the following equation:

$$d = \sum_{k=1}^{n}(|\,\bar{y}_k(\ddot{t})|)/n) \qquad (3)$$

Then the temperature changes rates deviation *E($\ddot{t}$)* will be calculated as follows:

$$E(\ddot{t}) = \bar{y}(\ddot{t})\text{-sign}(\bar{y}).d \qquad (4)$$

For ideal case, all values of *E($\ddot{t}$)* should be zero. But that doesn't happen in reality. The actual change rates deviate from the average value within certain displacement value *δ,* which depends on many factors such as: time slot width "s", the measurement place and the measurement period of the year. It can be determined by observation in normal measurement time periods. In case of air temperature, *δ* will be equal the modified standard deviation of *E($\ddot{t}$)*:

$$\delta = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(|\bar{y}_k(\ddot{t})| - d)^2} \qquad (5)$$

The values of $E(\ddot{t})$ will be categorized to three zones depending on δ according to the following equation classification:

$$\text{When} \begin{cases} |\mathrm{E}(\ddot{t})| \le \delta & : \text{Normal zone.} \\ \delta < |E(\ddot{t})| < 2\delta & : \text{Suspected zone.} \\ |\mathrm{E}(\ddot{t})| \ge 2\delta & : \text{Outlier zone .} \end{cases} \quad (6)$$

The following section will summarize the experimental result for the proposed algorithm.

2. Experimental evaluations

The proposed method has been applied to a randomly selected sample of temperature measured data in one day (20 June 2014) using DS18S20 sensors, which is a part of a full academic weather monitoring project. More details about the project can be found on the website “abc.altstu.ru”.

This sample has been divided into one hour time slot. Then the average for each slot has been calculated as well as $\bar{y}(\ddot{t})$, d, $\mathrm{E}\ (\ddot{t})$ and δ. The value of d and δ were 1.4 ^{o}C/H and 1.3 respectively.

In figure 3, measured temperature series, $\bar{y}(\ddot{t})$ and $\mathrm{E}(\ddot{t})$ have been plotted using Matlab.

As shown in the figure, the measured data has outlier region “the highlighted area”, which is outside the boundary 2δ.

To verify the results, the proposed algorithm has been applied to temperature sample from another weather station “at the city airport” measured at the same time period; the result, that shown in figure 2, indicates that the sample doesn’t contain any outliers, which in turn supports the validity of the method.

The cause of the outliers in the measured sample has been discovered. It was the effect of direct sunlight on the temperature sensor measurements.

3. Candlestick chart method

The same procedure can be applied using candlestick chart [19][20], but instead of change rate, the candle height H will be calculated as:

$$\mathrm{H}(\ddot{t}) = \mathrm{Close}((\ddot{t}) - \mathrm{Open}(\ddot{t}) \qquad 7)$$

Then,

$$\mathrm{d} = \textstyle\sum_{k=1}^{n}(|\ \mathrm{H_k}(\ddot{t})|)/\mathrm{n}) \qquad 8)$$

Where n is the number of candles.

$$\mathrm{E}(\ddot{t}) = \mathrm{H}(\ddot{t})\text{-sign}(\mathrm{H}(\ddot{t})).\mathrm{d} \qquad 9)$$

The value of δ will be calculated as in equation

This procedure has been also applied to the same sample of measured data; the result shown in figure 1 nearly is the same as in the change rate metho

4. Other outliers detection techniques

This section will review three of other well-known techniques in outliers and anomalies detection.

These techniques are modified Z-score, modified Thompson tau and modified boxplot.

- **Modified Z-score.** Z-scores are a very popular method for labeling outliers [21]. But the problem of Z-score is the effect of outlier on its calculations.

$$\mathrm{E(t)} = \frac{0.6745(\mathrm{f(t)} - \mathrm{f_m})}{\mathrm{MAD}} \qquad (10)$$

Where: f_m is the median value of f(t), and MAD = median(|f(t)-f_m |)

The authors recommend that modified Z-scores with an absolute value of $E(t)$ greater than the threshold value δ =3.5 will be considered as potential outliers.

This technique has be applied to the measured sample, but it failed to detect the outliers and all the absolute values of E(t) were less than the threshold value 3.5. However, a small modification for calculation of the threshold value δ maybe solves the problem. The following equation presents a proposed value for δ:

$$\delta = \mathrm{median}(|\ |\mathrm{f(t)}\text{-}\mathrm{f_m}\ |\text{-}\mathrm{MAD}|) \qquad (11)$$

- **The modified Thompson tau technique.** The Thompson tau technique is excellent for rejecting outliers, but also may reject some good data, so it is better to use the modified Thompson tau technique [22]. This method takes into account a data set’s standard deviation, average and provides a statistically determined rejection zone; thus providing an objective method to determine outliers. It will be summarized in the following steps:

- The sample mean $\bar{f}$ and the sample standard deviation S_f are calculated as usual.
- For each data point, the absolute value of the deviation will calculated as:

$$\tilde{\mathrm{f}}(\mathrm{t}) = |\mathrm{f(t)}\text{-}\ \bar{\mathrm{f}}| \qquad (12)$$

- The value of the modified Thompson **τ** is calculated from the following equation:

$$\boldsymbol{\tau} = \frac{\mathbf{t}_{\alpha/2}\cdot(\mathbf{N-1})}{\sqrt{N}\ \sqrt{N-2+t^2{}_{\alpha/2}}} \qquad (13)$$

Where: N is the number of the sample points, $t_{\alpha/2}$ is the critical student’s t value. “It can be calculated using Matlab built-in function TINV”.

- Then the outliers can be detected using the following classification:
  - If $\tilde{\boldsymbol{f}}(\boldsymbol{t})$ **> τ* S_f ,** the sample point is outlier.
  - If $\tilde{\boldsymbol{f}}(\boldsymbol{t}) \le$ **τ* S_f ,** the sample point is not outlier.

This technique has been applied to the measured sample and the result is shown in figure 5.

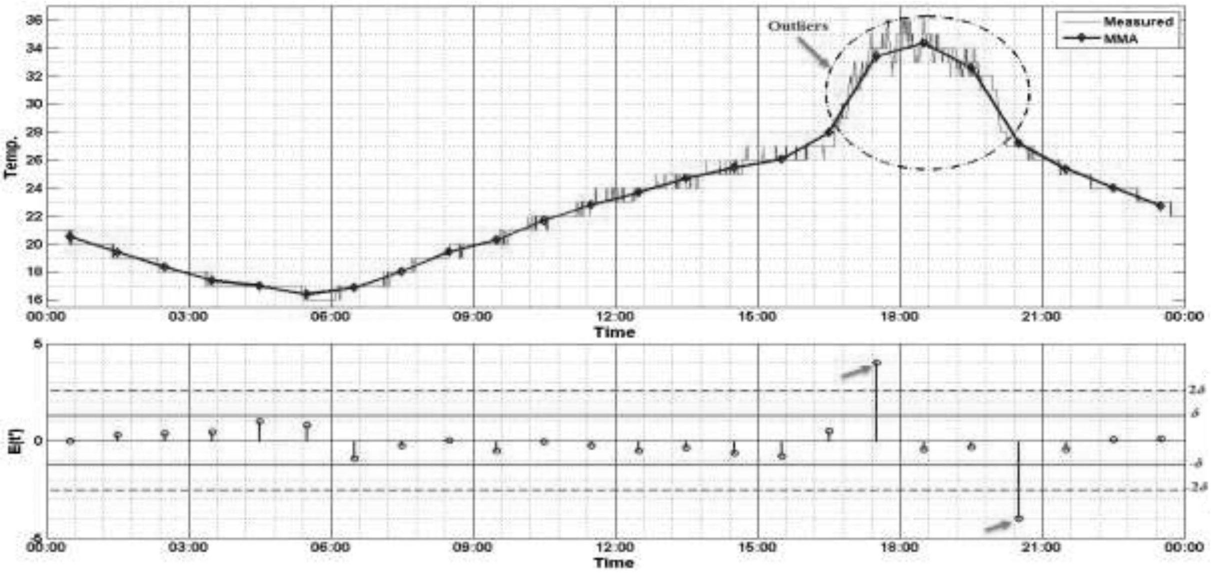

Fig. 3. Outliers detection using the change rate method.

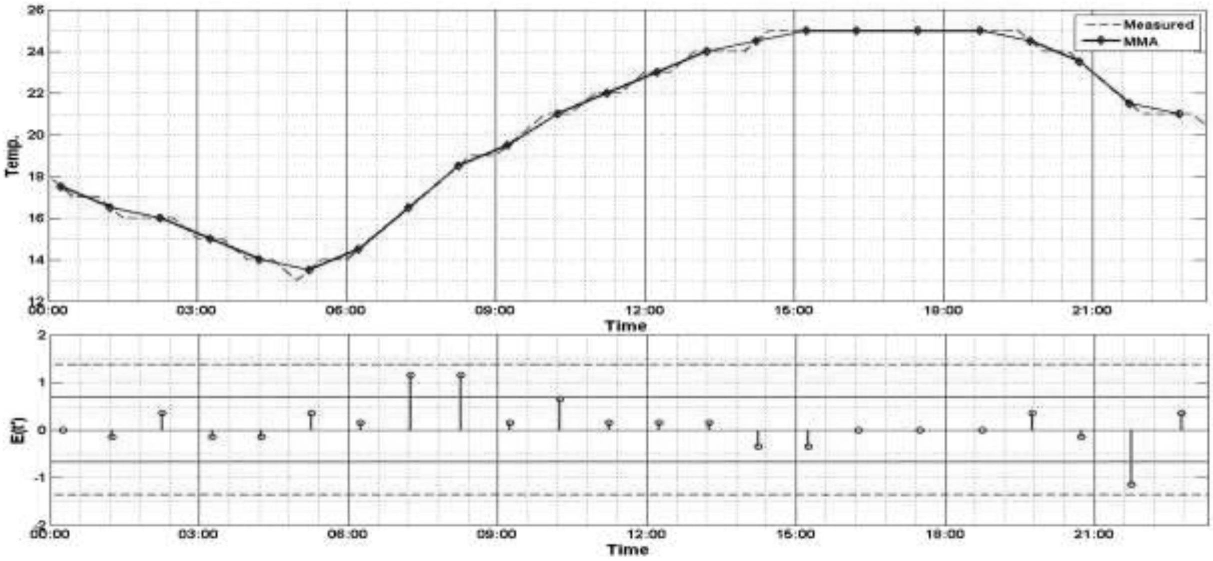

Fig. 2. Termperature sample in 20/6/2014 at Barnaul airport.

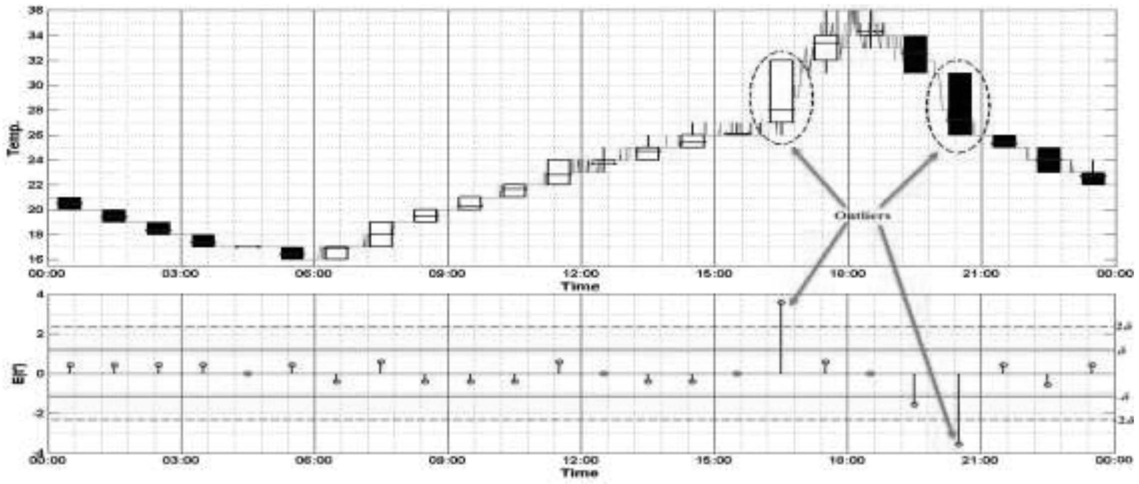

Fig. 1. Outliers detection using candlestick chart.

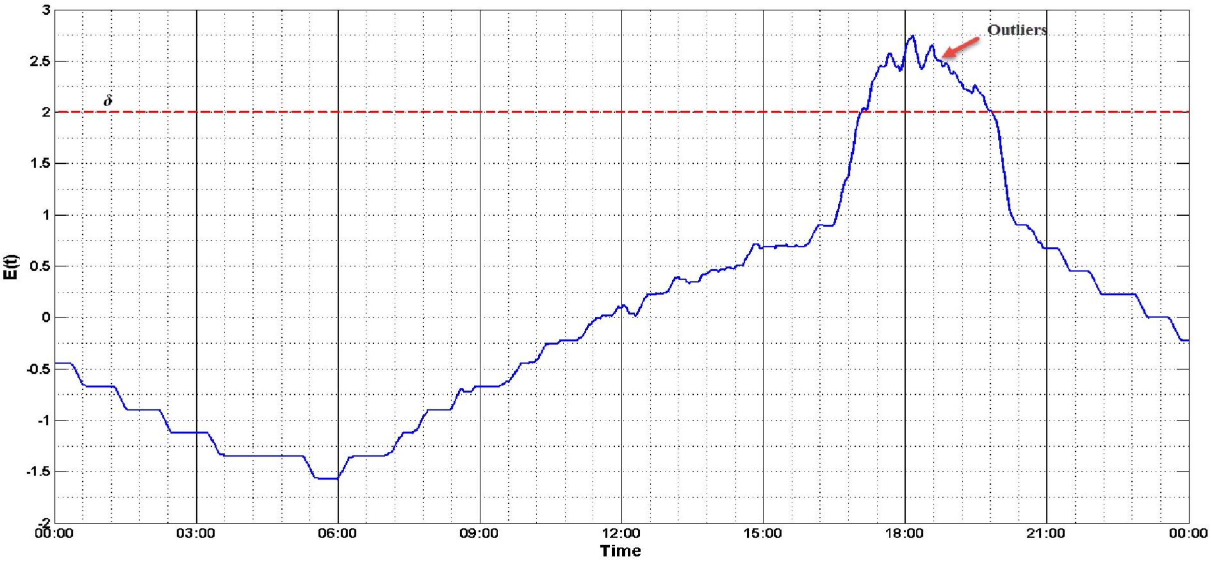

Fig. 3. Outliers detection using modified Z-score technique.

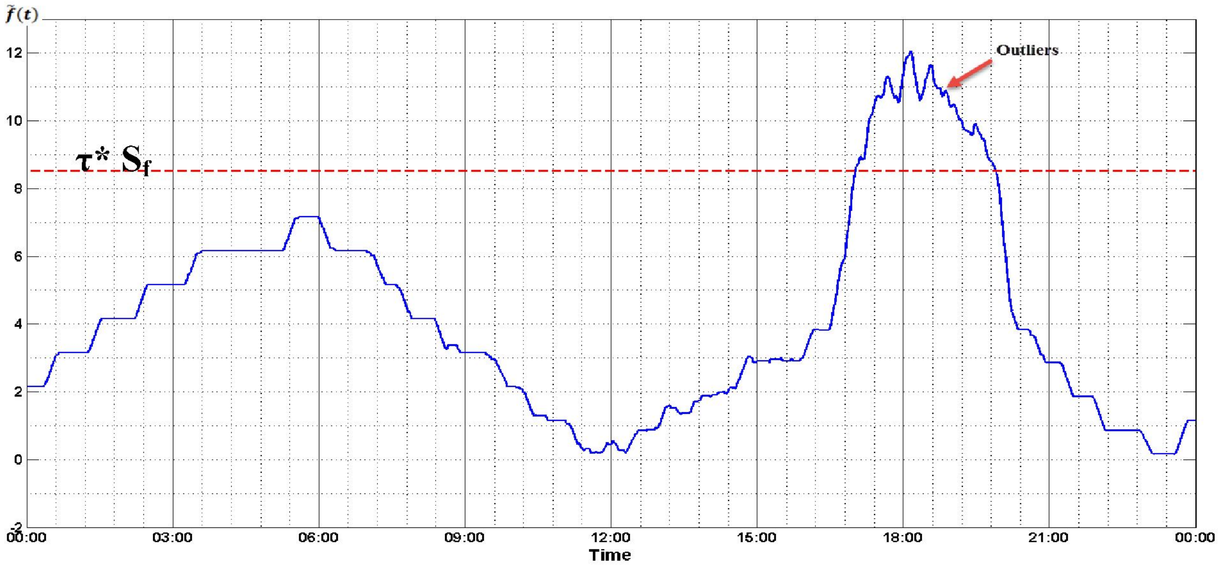

Fig. 2. Outliers detection using The modified Thompson tau technique.

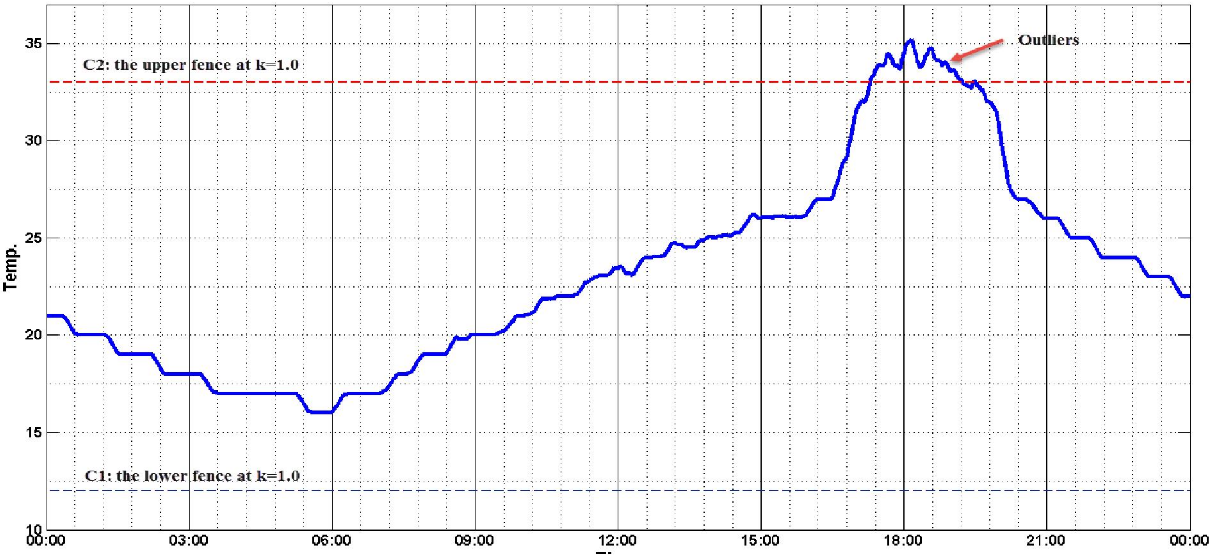

Fig. 1. Outliers detection using modified Boxplot.

- **The adjusted boxplot.** Boxplots [23] display variation in data samples without making any assumptions of the underlying statistical distribution: boxplots are non-parametric. The spacings between the different parts of the box indicate the degree of spread and data skewness, and show outliers. However, this method has a limitation in outliers' detection, especially for highly skewed measurements.

The adjusted boxplot [24] considers the medcouple (MC), a robust measure of skewness for a skewed distribution.

MC is defined as [25]:

$$\mathrm{MC} = \mathrm{Median}_{f_i \leq f_m \leq f_j} h(f_i, f_j) \quad (1)$$

With f_m the sample median and where for all $f_i \neq f_j$ the kernel "h" function is given by:

$$h(f_i, f_j) = \frac{f_j - f_m \ -(f_m - f_i)}{f_j - f_i} \quad (2)$$

The medcouple always lies between -1 and 1. A distribution that is skewed to the right has a positive value for the medcouple, whereas it becomes negative at a left skewed distribution. Finally, a symmetric distribution has a zero medcouple.

According to [26] the interval of the adjusted boxplot is:

$$\begin{aligned} C_1 &= \begin{matrix} Q1 - ke^{-3.5MC} \ \ Q3 - Q1 \ , & \text{if } MC \geq 0 \\ Q1 - ke^{-4MC} \ \ Q3 - Q1 \ , & \text{if } MC \leq 0 \end{matrix} \\ C_2 &= \begin{matrix} Q3 + ke^{4MC} \ \ Q3 - Q1 \ , & \text{if } MC \geq 0 \\ Q3 + ke^{3.5MC} \ \ Q3 - Q1 \ , & \text{if } MC \leq 0 \end{matrix} \end{aligned} \quad (3)$$

Where C_1 is the lower fence and C_2 is the upper fence of the interval. The observations which fall outside the interval are considered outliers. The author in [23] has suggested k=1.5 for the lower fence and k=3.0 for the upper fence. Whereas [27] used k=1.0 and k=1.5. But [28] used k=2. So, which of these values should be used? The authors in [29] answered the question and preferred the standard value for k=1.5.

But, when this methods has been applies on the selected measured sample with k=1.5, it failed to detect the outliers. The fences have been recalculated with k=1.0, it almost detected the top portion of the outlier. The simulation result is shown in figure 6.

1. Result comparison and dissection

In Previous sections, two new outliers methods detection were proposed and simulated. Also, three general techniques were reviewed and applied for measured sample.

The change rate detected the outliers' boundaries accurately. The accuracy comes from the calculation way of the fence "δ", which depends on change rate variation and standard deviation.

Another advantage of that technique was identifying a suspicious zone. A farther study for this zone of data will give very variable information like novelty.

References

1. Grubbs F.E. Procedures for Detecting Outlying Observations in Samples // Technometrics. 1969. Vol. 11. P. 1–21.
2. Barnett V., Lewis T. Outliers in Statistical Data. 3th ed. Wiley, 1994.
3. Ben-gal I. Outlier Detection // Data Mining and Knowledge Discovery Handbook. 2005. P. 131–146.
4. Aggarwal C.C., Zhao Y., Yu P.S. Outlier detection in graph streams // Proceedings - International Conference on Data Engineering. 2011. P. 399–409.
5. Barnett V. The Study of Outliers: Purpose and Model // J. R. Stat. Soc. Ser. C (Applied Stat. 1978. Vol. 27. P. 242–250.
6. Horn P.S. et al. Effect of outliers and nonhealthy individuals on reference interval estimation. // Clin. Chem. 2001. Vol. 47. P. 2137–2145.
7. Solberg H.E., Lahti A. Detection of outliers in reference distributions: performance of Horn's algorithm. // Clin. Chem. 2005. Vol. 51. P. 2326–2332.
8. Clifton D.A., Hugueny S., Tarassenko L. Novelty detection with multivariate extreme value statistics // J. Signal Process. Syst. 2011. Vol. 65. P. 371–389.
9. Keogh E., Lonardi S., Chiu B. "Yuan-chi." Finding surprising patterns in a time series database in linear time and space // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02. 2002. P. 550.
10. Lu T.C., Juang J.C., Yu G.R. On-line outliers detection by neural network with quantum evolutionary algorithm // Second International Conference on Innovative Computing, Information and Control, ICICIC 2007. 2008.
11. Bakar Z. et al. A Comparative Study for Outlier Detection Techniques in Data Mining // 2006 IEEE Conf. Cybern. Intell. Syst. 2006. P. 1–6.
12. Sane S.S., Ghatol A.A. Use of instance typicality for efficient detection of outliers with neural network classifiers // Proceedings - 9th International Conference on Information Technology, ICIT 2006. 2007. P. 225–228.
13. Chatzigiannakis V. et al. Hierarchical anomaly detection in distributed large-scale sensor networks // Proceedings - International Symposium on Computers and Communications. 2006. P. 761–766.
14. Subramaniam S. et al. Online outlier detection in sensor data using non-parametric models // VLDB '06 Proc. 32nd Int. Conf. Very large data bases. 2006. P. 187–198.
15. Idé T., Papadimitriou S., Vlachos M. Computing correlation anomaly scores using stochastic nearest neighbors // Proceedings - IEEE International Conference on Data Mining, ICDM. 2007. P. 523–528.
16. Albrecht S. et al. Generalized radial basis function networks for classification and novelty detection: Self-organization of optimal Bayesian decision // Neural Networks. 2000. Vol. 13. P. 1075–1093.

17. Janakiram D. et al. Outlier Detection in Wireless Sensor Networks using Bayesian Belief Networks // 2006 1st International Conference on Communication Systems Software & Middleware. 2006. P. 1–6.
18. Ya-Lun C. Statistical Analysis. 2nd ed. Holt,Rinehart & Winston of Canada Ltd, 1975. 894 p.
19. Rhoads R. Candlestick Charting For Dummies. John Wiley & Sons, 2011. 360 p.
20. Person J.L. Candlestick and Pivot Point Trading Triggers: Setups for Stock, Forex, and Futures Markets. John Wiley & Sons, 2011. 368 p.
21. Shiffler R.E. Maximum Z Scores and Outliers // Am. Stat. 1988. Vol. 42. P. 79–80.
22. Thompson W.R. On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation // Ann. Math. Stat. Institute of Mathematical Statistics, 1935. Vol. 6, № 4. P. 214–219.
23. Tukey J.W. Exploratory Data Analysis // Analysis / ed. Wrigley N., Bennet R.J. Addison-Wesley, 1977. Vol. 2, № 1999. 688 p.
24. Hubert M., Vandervieren E. An adjusted boxplot for skewed distributions // Comput. Stat. Data Anal. 2008. Vol. 52. P. 5186–5201.
25. Brys G., Hubert M., Struyf A. A Robust Measure of Skewness // J. Comput. Graph. Stat. Taylor & Francis, 2004. Vol. 13, № 4. P. 996–1017.
26. Brys G., Hubert M., Rousseeuw P.J. A robustification of independent component analysis // J. Chemom. 2005. Vol. 19, № 5-7. P. 364–375.
27. McNeil D.R. Interactive data analysis: a practical primer. John Wiley & Sons Australia, Limited, 1977. 186 p.
28. Ingelfinger J.A. Biostatistics in clinical medicine. Macmillan, 1983. 316 p.
29. Frigge M., Hoaglin D.C., Iglewicz B. Some Implementations of the Boxplot // Am. Stat. 1989. Vol. 43. P. 50–54.

Аспирант Хуссейн Х.М.- *(Египет, E-mail: helphs@yahoo.com) и*

Якунин А.Г. *- д.т.н., профессор E-mail: yakunin@agtu.secna.ru) - кафедра вычислительных систем и информационной безопасности ФГБОУ ВПО «Алтайский государственный технический университет им. И.И. Ползунова», г. Барнаул*